SATHYABAMA
INSTITUTE OF SCIENCE AND TECHNOLOGY
(DEEMED TO BE UNIVERSITY)
Accredited "A" Grade by NAAC | 12B Status by UGC | Approved by AICTE
www.sathyabama.ac.in

SCHOOL OF BIO AND CHEMICAL ENGINEERING

DEPARTMENT OF BIOINFORMATICS

UNIT – 1- SBIA5304 MICROARRAY DATAANALYSIS

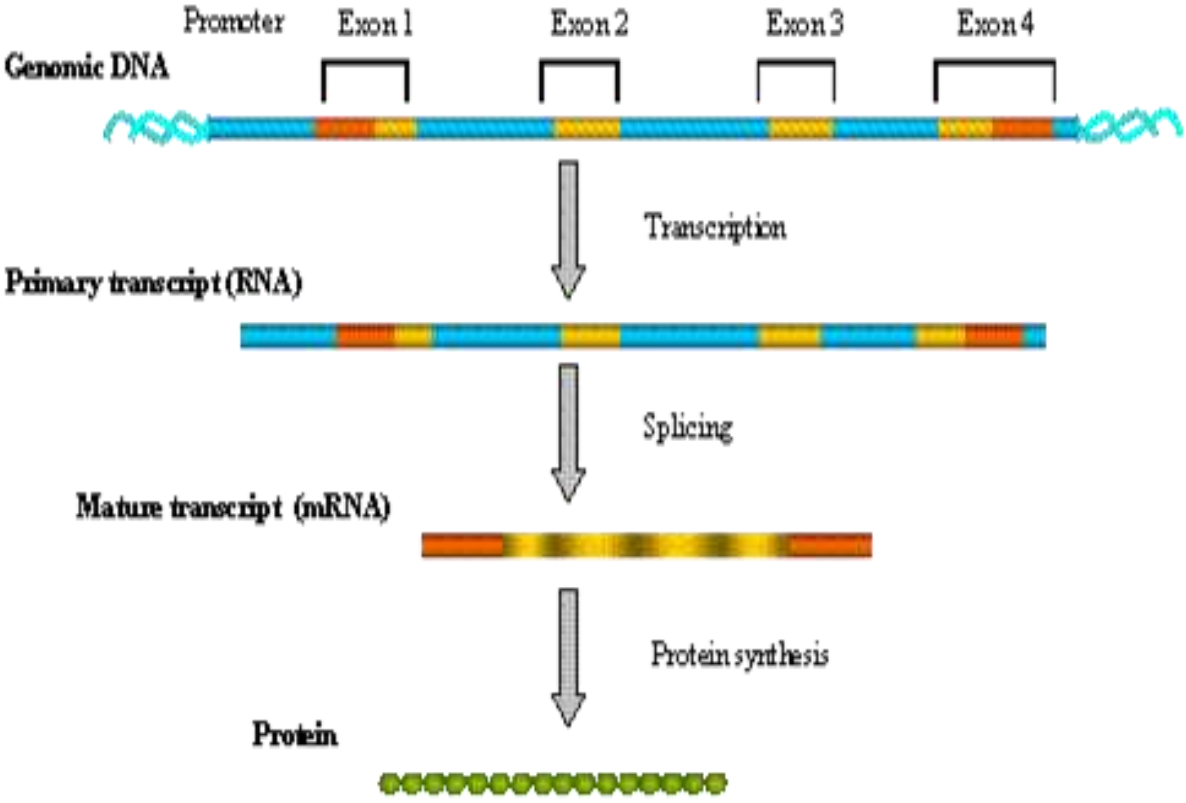
SBIA5304- MICROARRAY DATA ANALYSIS

UNIT I GENE EXPRESSION

Basics of Gene expression- definition , gene expression studies, gene expression patterns- Applications of gene expression studies

Microarrays – definition , discovery, technique, making microarrays, spotted microarrays, In-situ synthesized oligonucleotide arrays, inkjet array synthesis, Affymetrix techniques, DNA CHIP technology, photolithography, spot quality, sample preparation and labelling, washing, image acquisition Sequencing by Hybridization Arrays- DNA MassArray™ Technology- Printing DNA Microarrays-Types of microarrays - Designing a microarray experiment.

- Gene, unit of hereditary information that occupies a fixed position (locus) on a chromosome. Genes achieve their effects by directing the synthesis of proteins.
- Gene, unit of hereditary information that occupies a fixed position (locus) on a chromosome. Genes achieve their effects by directing the synthesis of proteins.



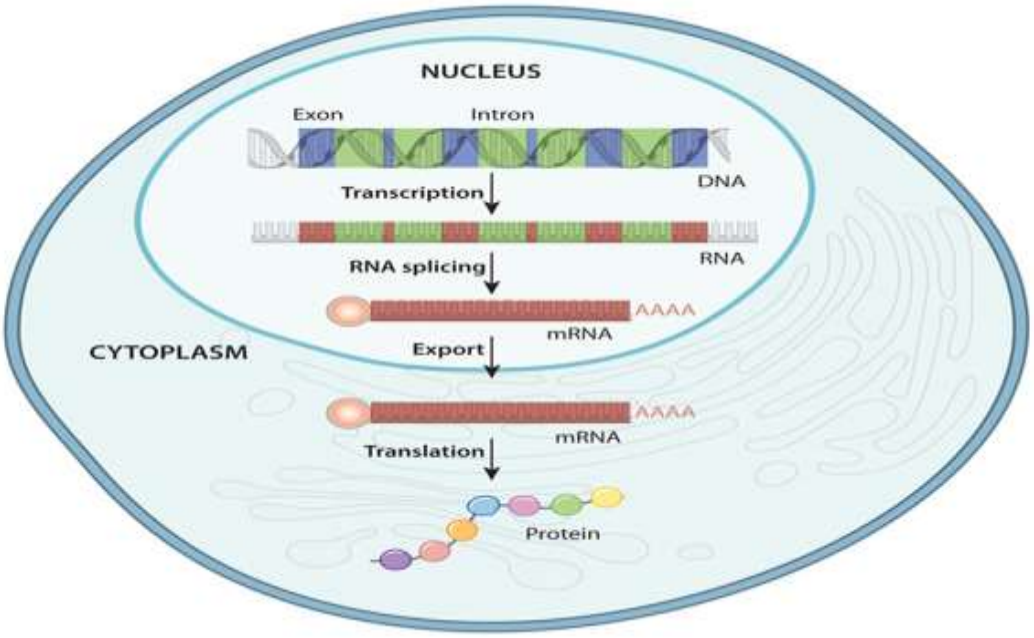
- In eukaryotes (such as animals, plants, and fungi), genes are contained within the cell nucleus.

- The mitochondria (in animals) and the chloroplasts (in plants) also contain small subsets of genes distinct from the genes found in the nucleus.
- In prokaryotes (organisms lacking a distinct nucleus, such as bacteria), genes are contained in a single chromosome that is free-floating in the cell cytoplasm.
- Many bacteria also contain plasmids—extrachromosomal genetic elements with a small number of genes.
- The number of genes in an organism's genome (the entire set of chromosomes) varies significantly between species.
- For example, whereas the human genome contains an estimated 20,000 to 25,000 genes, the genome of the bacterium *Escherichia coli O157:H7* houses precisely 5,416 genes.
- *Arabidopsis thaliana*—the first plant for which a complete genomic sequence was recovered—has roughly 25,500 genes; its genome is one of the smallest known to plants.
- Among extant independently replicating organisms, the bacterium *Mycoplasma genitalium has the fewest number of genes, just 517.*
- Basics of Gene expression:
- Gene expression
- the phenotypic manifestation of a gene or genes by the processes of genetic transcription and genetic translation.
- Gene expression analysis
- the determination of the pattern of genes expressed at the level of genetic transcription, under specific circumstances or in a specific cell.
- When genes are expressed, the genetic information (base sequence) on DNA is first copied to a molecule of mRNA (transcription).
- The mRNA molecules then leave the cell nucleus and enter the cytoplasm, where they participate in protein synthesis by specifying the particular amino acids that make up individual proteins (translation).
- At any given time, the amount of a particular protein in a cell reflects the balance between that protein's synthetic and degradative biochemical pathways.

- On the synthetic side of this balance, recall that protein production starts at transcription (DNA to RNA) and continues with translation (RNA to protein).
- Thus, control of these processes plays a critical role in determining what proteins are present in a cell and in what amounts.
- In addition, the way in which a cell processes its RNA transcripts and newly made proteins also greatly influences protein levels.

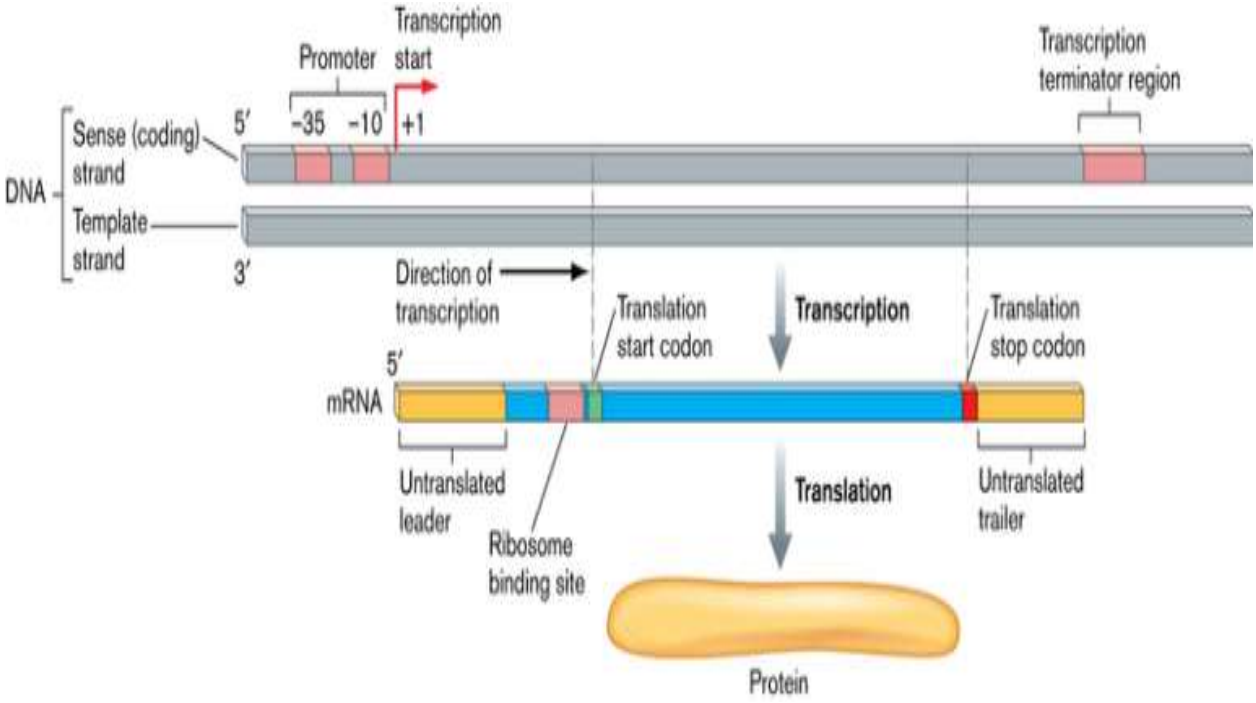
How Is Gene Expression Regulated?

- The amounts and types of mRNA molecules in a cell reflect the function of that cell. In fact, thousands of transcripts are produced every second in every cell. Given this statistic, it is not surprising that the primary control point for gene expression is usually at the very beginning of the protein production process — the initiation of transcription. RNA transcription makes an efficient control point because many proteins can be made from a single mRNA molecule.

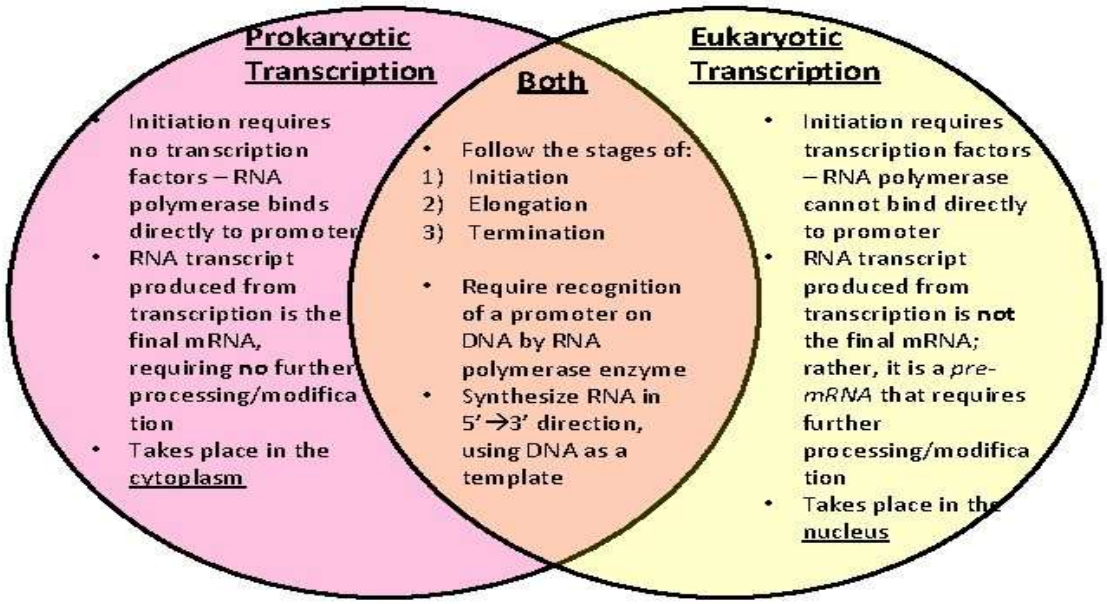


Transcript processing provides an additional level of regulation for eukaryotes, and the presence of a nucleus makes this possible.

- **In prokaryotes**, translation of a transcript begins before the transcript is complete, due to the proximity of ribosomes to the new mRNA molecules.

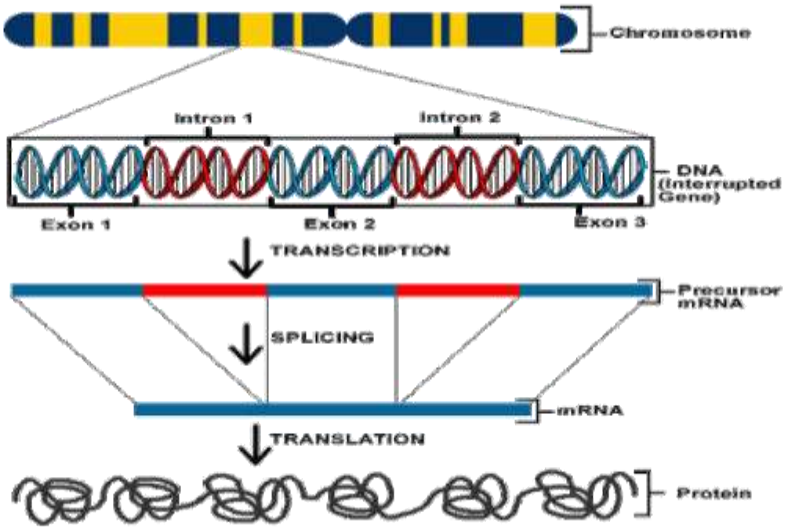


- In eukaryotes, however, transcripts are modified in the nucleus before they are exported to the cytoplasm for translation.



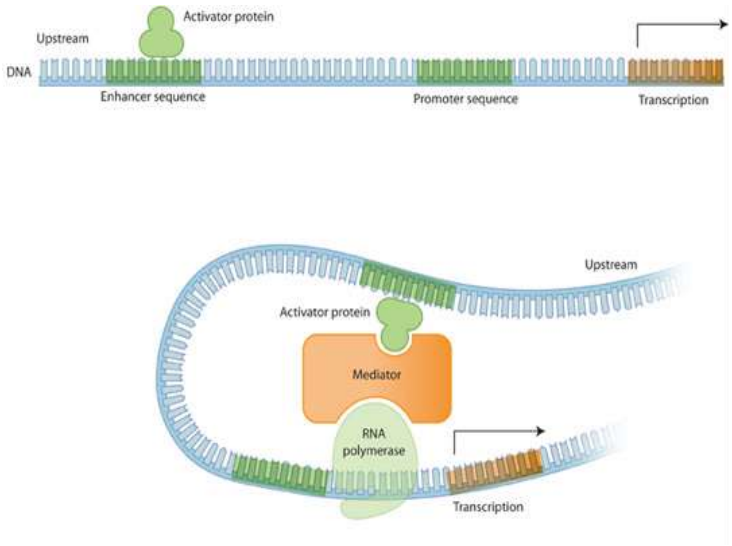
- Eukaryotic transcripts are also more complex than prokaryotic transcripts.

- For instance, the primary transcripts synthesized by RNA polymerase contain sequences that will not be part of the mature RNA.
- These intervening sequences are called introns, and they are removed before the mature mRNA leaves the nucleus.
- The remaining regions of the transcript, which include the protein-coding regions, are called exons, and they are spliced together to produce the mature mRNA.
- Eukaryotic transcripts are also modified at their ends, which affects their stability and translation.
- Of course, there are many cases in which cells must respond quickly to changing environmental conditions.
- In these situations, the regulatory control point may come well after transcription.
- For example, early development in most animals relies on translational control because very little transcription occurs during the first few cell divisions after fertilization. Eggs therefore contain many maternally originated mRNA transcripts as a ready reserve for translation after fertilization.
- On the degradative side of the balance, cells can rapidly adjust their protein levels through the enzymatic breakdown of RNA transcripts and existing protein molecules.
- Both of these actions result in decreased amounts of certain proteins. Often, this breakdown is linked to specific events in the cell.
- The eukaryotic cell cycle provides a good example of how protein breakdown is linked to cellular events. This cycle is divided into several phases, each of which is characterized by distinct cyclin proteins that act as key regulators for that phase.
- Before a cell can progress from one phase of the cell cycle to the next, it must degrade the cyclin that characterizes that particular phase of the cycle. Failure to degrade a cyclin stops the cycle from continuing.



How Do Different Cells Express the Genes They Need?

- Only a fraction of the genes in a cell are expressed at any one time.
- The variety of gene expression profiles characteristic of different cell types arise because these cells have distinct sets of transcription regulators.
- Some of these regulators work to increase transcription, whereas others prevent or suppress it.
- Normally, transcription begins when an RNA polymerase binds to a so-called promoter sequence on the DNA molecule.
- This sequence is almost always located just upstream from the starting point for transcription (the 5' end of the DNA), though it can be located downstream of the mRNA (3' end).
- In recent years, researchers have discovered that other DNA sequences, known as enhancer sequences, also play an important part in transcription by providing binding sites for regulatory proteins that affect RNA polymerase activity.
- Binding of regulatory proteins to an enhancer sequence causes a shift in chromatin structure that either promotes or inhibits RNA polymerase and transcription factor binding. A more open chromatin structure is associated with active gene transcription.
- In contrast, a more compact chromatin structure is associated with transcriptional *in*activity

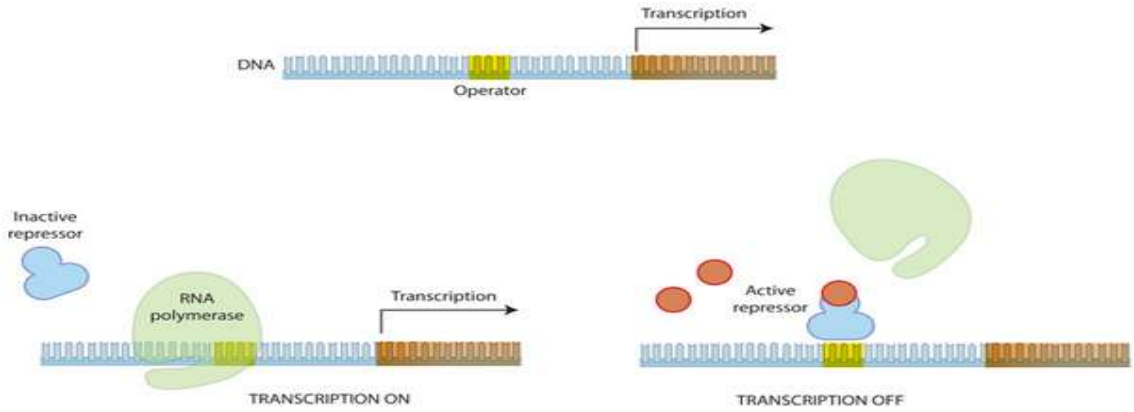


- Some regulatory proteins affect the transcription of multiple genes.
- This occurs because multiple copies of the regulatory protein binding sites exist within the genome of a cell.
- Consequently, regulatory proteins can have different roles for different genes, and this is one mechanism by which cells can coordinate the regulation of many genes at once.

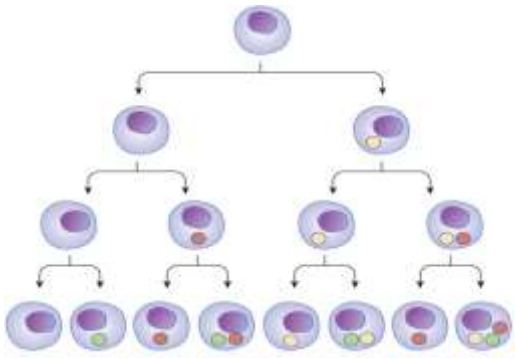
How Is Gene Expression Increased or Decreased in Response to Environmental Change?

- In prokaryotes, regulatory proteins are often controlled by nutrient availability.
- This allows organisms such as bacteria to rapidly adjust their transcription patterns in response to environmental conditions.
- In addition, regulatory sites on prokaryotic DNA are typically located close to transcription promoter sites — and this plays an important part in gene expression.
- For an example of how this works, imagine a bacterium with a surplus of amino acids that signal the turning "on" of some genes and the turning "off" of others.
- In this particular example, cells might want to turn "on" genes for proteins that metabolize amino acids and turn "off" genes for proteins that synthesize amino acids.
- Some of these amino acids would bind to positive regulatory proteins called activators.

- Activator proteins bind to regulatory sites on DNA nearby to promoter regions that act as on/off switches. This binding facilitates RNA polymerase activity and transcription of **nearby genes.**



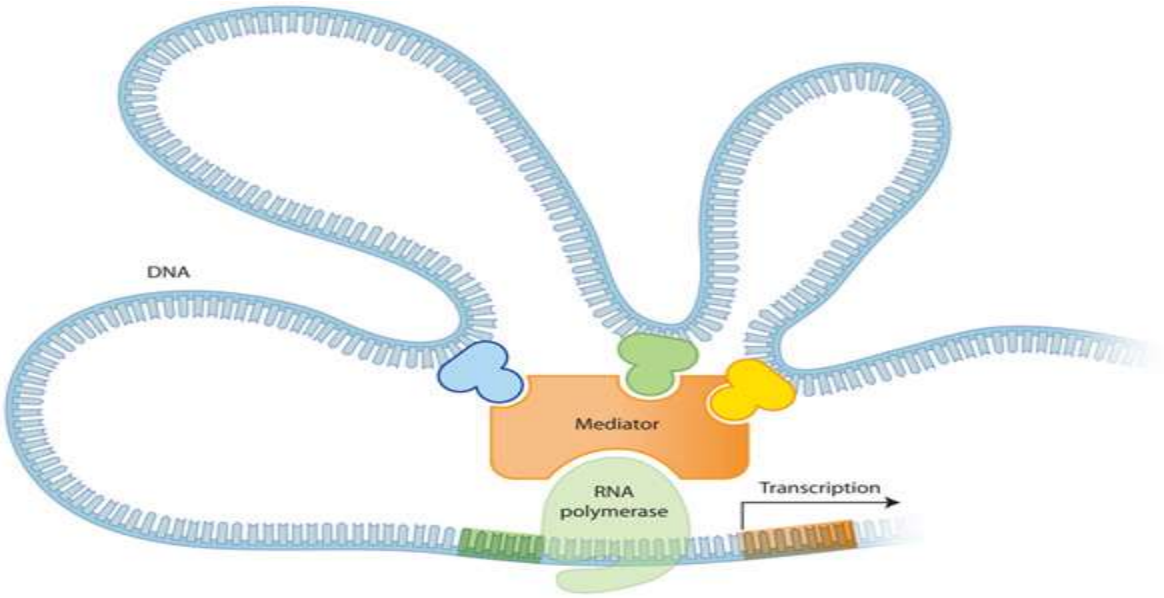
- At the same time, however, other amino acids would bind to negative regulatory proteins called repressors, which in turn bind to regulatory sites in the DNA that effectively block RNA polymerase binding
- The control of gene expression in eukaryotes is more complex than that in prokaryotes. In general, a greater number of regulatory proteins are involved, and regulatory binding sites may be located quite far from transcription promoter sites. Also, eukaryotic gene expression is usually regulated by a combination of several regulatory proteins acting together, which allows for greater flexibility in the control of gene expression.
- Different cell types express characteristic sets of transcriptional regulators. In fact, as multicellular organisms develop, different sets of cells within these organisms turn specific combinations of regulators on and off. Such developmental patterns are responsible for the variety of cell types present in the mature organism



- ***Transcriptional regulators can determine cell types***

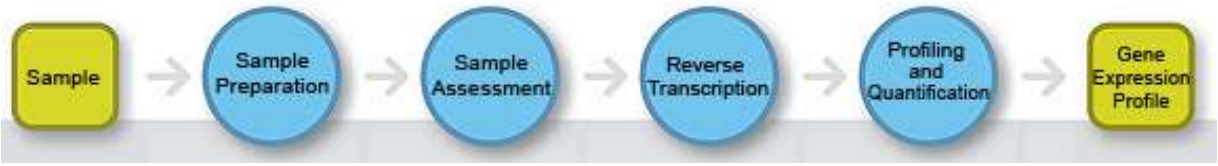
- *The wide variety of cell types in a single organism can depend on different transcription factor activity in each cell type. Different transcription factors can turn on at different times during successive generations of cells. As cells mature and go through different stages (arrows), transcription factors (colored balls) can act on gene expression and change the cell in different ways. This change affects the next generation of cells derived from that cell. In subsequent generations, it is the combination of different transcription factors that can ultimately determine cell type.*
- Technologies
- Real Time quantitative RT-PCR
- In situ hybridization
- Microarrays
- Massively Parallel Signature Sequencing (MPSS)
- **Gene expression analysis** is most simply described as the study of the way genes are transcribed to synthesize functional gene products — functional RNA species or protein products.
- The study of gene regulation provides insights into normal cellular processes, such as differentiation, and abnormal or pathological processes.
- In 1941, Beadle and Tatum published experiments that would explain the basis of the central dogma of molecular biology, whereby the DNA through an intermediate molecule, called RNA, results proteins that perform the functions in cells.
- Currently, biomedical research attempts to explain the mechanisms by which develops a particular disease, for this reason, gene expression studies have proven to be a great resource.
- Strictly, the term “gene expression” comprises from the gene activation until the mature protein is located in its corresponding compartment to perform its function and contribute to the expression of the phenotype of cell.
- The expression **studies are directed to detect and quantify messenger RNA (mRNA) levels of a specific gene.**
- The **development of the RNA-based gene expression studies began with the Northern Blot by Alwine et al. in 1977.**

- In **1969, Gall and Pardue and John et al. independently developed the in situ hybridization, but this technique was not employed to detect mRNA until 1986 by Coghlan**. Today, many of the techniques for quantification of RNA are deprecated because other new techniques provide more information. Currently the most widely used techniques are qPCR, expression microarrays, and RNAseq for the transcriptome analysis. In this chapter, these techniques will be reviewed.
- **Gene expression workflow.**
- Researchers may perform gene expression analysis at any one of several different levels at which gene expression is regulated: transcriptional, post-transcriptional, translational, and post-translational
- protein modification.
- Transcription, the process of creating a complementary RNA copy of a DNA sequence, can be regulated in a variety of ways. Transcriptional regulation processes are the most commonly studied and manipulated in typical gene expression analysis experiments.
-



-

As previously mentioned, enhancer sequences are DNA sequences that are bound by an activator protein, and they can be located thousands of base pairs away from a promoter, either upstream or downstream from a gene. Activator protein binding is thought to cause DNA to loop out, bringing the activator protein into physical proximity with RNA polymerase and the other proteins in the complex that promote the initiation of transcription

- 

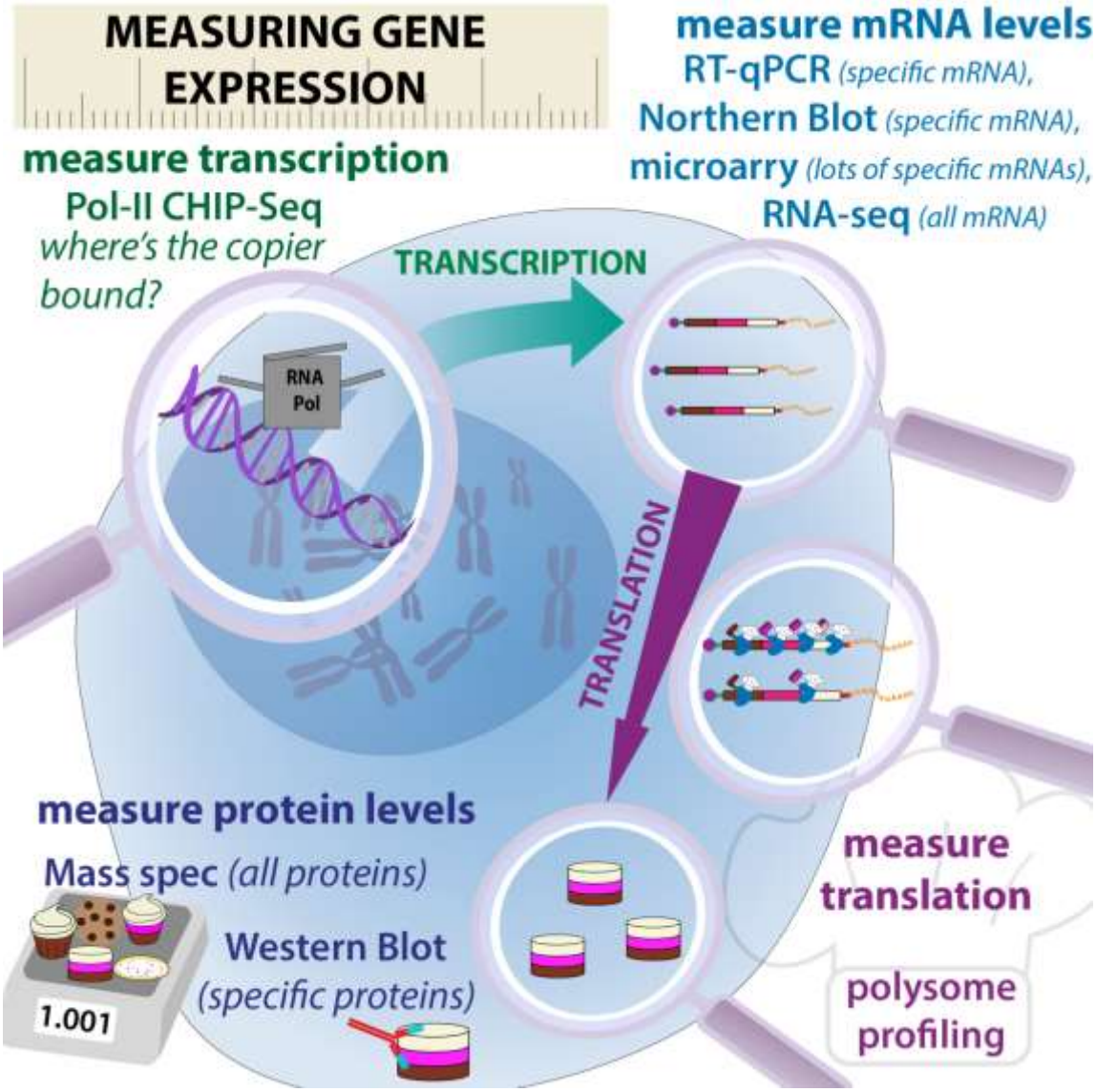
  The binding of regulatory proteins to DNA binding sites is the most direct method by which transcription is naturally modulated. Alternatively, regulatory processes can also interact with the transcriptional machinery of a cell. More recently, the influence of epigenetic regulation, such as the effect of variable DNA methylation on gene expression, has been uncovered as a powerful tool for gene expression profiling. Varying degrees of methylation are known to affect chromatin folding and strongly affect accessibility of genes to active transcription.

- Following transcription, eukaryotic RNA is typically spliced to remove noncoding intron sequences and capped with a poly(A) tail. At this post-transcriptional level, RNA stability has a significant effect on functional gene expression, that is, the production of functional protein. Small interfering RNA (siRNA) consists of double-stranded nucleic acid molecules that are participants in the RNA interference pathway, in which the expression of specific genes is modulated (typically by decreasing activity). Precisely how this modulation is accomplished is not yet fully understood. A growing field of gene expression analysis is in the area of microRNAs (miRNAs), short RNA molecules that also act as eukaryotic post-transcriptional regulators and gene silencing agents

Researchers studying gene expression employ a wide variety of molecular biology techniques and experimental methods.

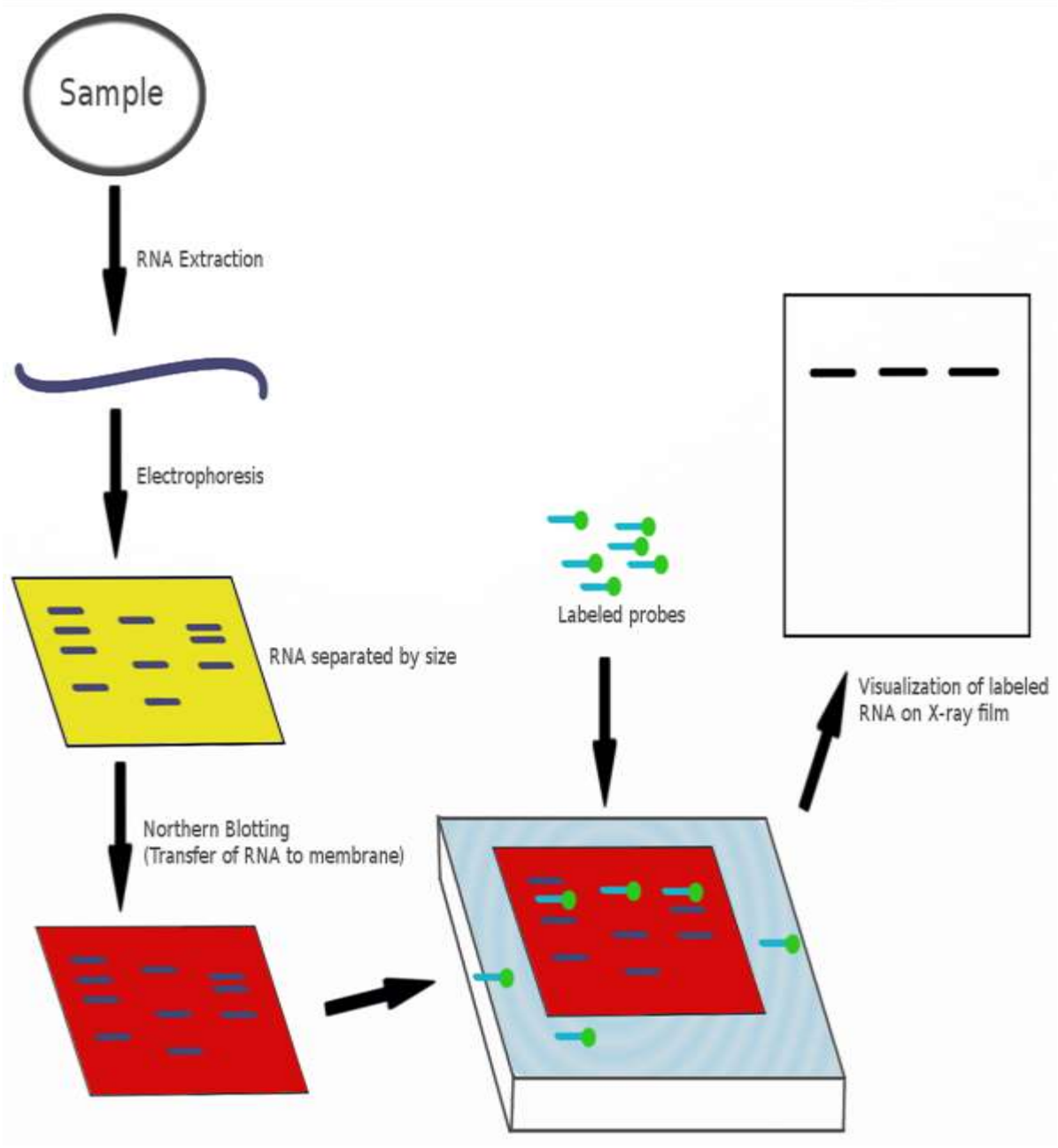
Gene expression analysis studies can be broadly divided into four areas:

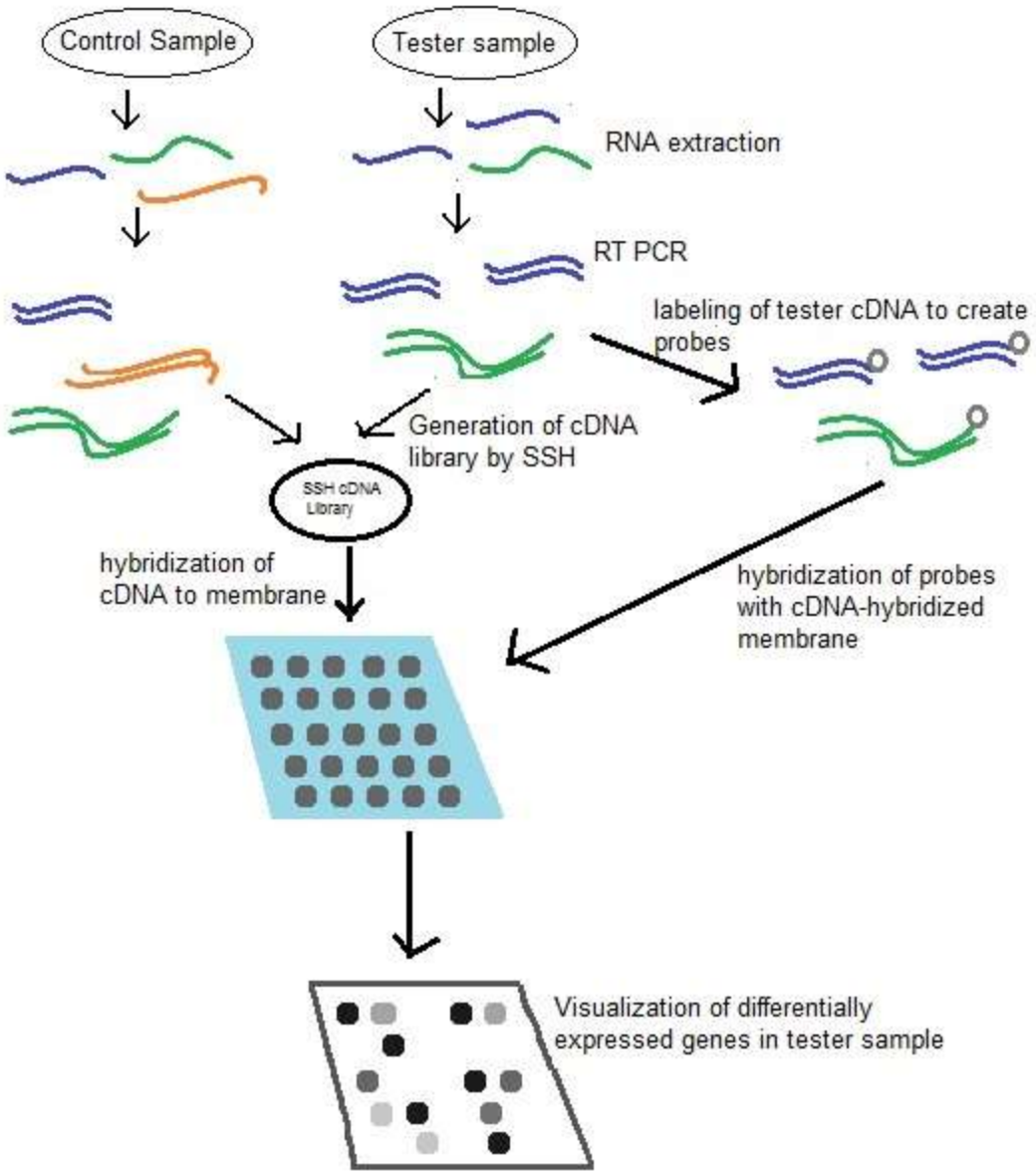
1. **RNA expression,**
2. **promoter analysis,**
3. **protein expression, and**
4. **post-translational modification.**



RNA Expression

- Northern blotting — steady-state levels of mRNA are directly quantitated by electrophoresis and transfer to a membrane followed by incubation with specific probes.
- The RNA-probe complexes can be detected using a variety of different chemistries or radionuclide labeling.
- This relatively laborious technique was the first tool used to measure RNA levels


Sample
RNA Extraction
Electrophoresis
RNA separated by size
Northern Blotting
(Transfer of RNA to membrane)
Labeled probes
Visualization of labeled
RNA on X-ray film



The reverse northern blot is a method by which gene expression patterns may be analyzed by comparing isolated RNA molecules from a tester sample to samples in a control cDNA library.

- It is a variant of the northern blot in which the nucleic acid immobilized on a membrane is a collection of isolated DNA fragments rather than RNA, and the probe is RNA extracted from a tissue and radioactively labelled.

- A reverse northern blot can be used to profile expression levels of particular sets of RNA sequences in a tissue or to determine presence of a particular RNA sequence in a sample.
- Although DNA Microarrays and newer next-generation techniques have generally supplanted reverse northern blotting, it is still utilized today and provides a relatively cheap and easy means of defining expression of large sets of genes.

Procedure

- In order to prepare the reverse northern membrane, cDNA sequences for transcripts of interest are immobilized on nylon membranes, which can be accomplished by use of dot blots or bidirectional agarose gel blotting and UV fixation of the DNA to the membranes.
- In many cases, cDNA probes may be preferred over RNA probes in order to mitigate problems of RNA degradation by RNAses or tissue metabolites.
- Prepared reverse northern blot membranes are pre-hybridized in Denhardt's solution with SSC buffer and labeled cDNA probes are denatured at 100 °C and added to the pre-hybridization solution.
- The membrane is incubated with the probes for at least 15 hours at 65 °C, then washed and exposed.

DNA microarrays:

- an array of oligonucleotide probes bound to a chip surface enables gene expression profiling of many genes in response to a condition.
- Labeled cDNA from a sample is hybridized to complementary probe sequences on the chip, and strongly associated complexes are identified optically.
- Gene expression profiling is often a first step in a gene expression analysis workflow, investigating changes in the expression profile of a whole system or examining the effects of mutations in biological systems

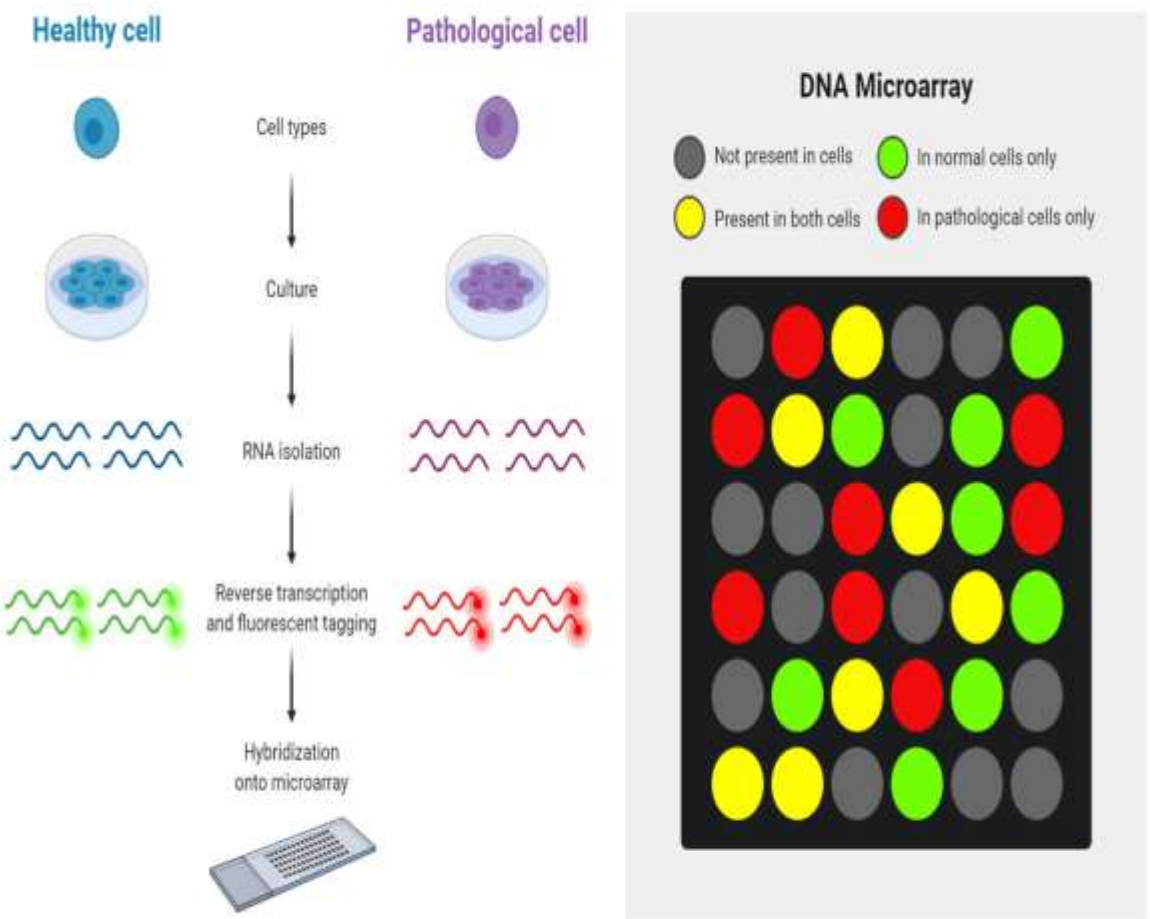


Image By Sagar Aryal, created using biorender.com

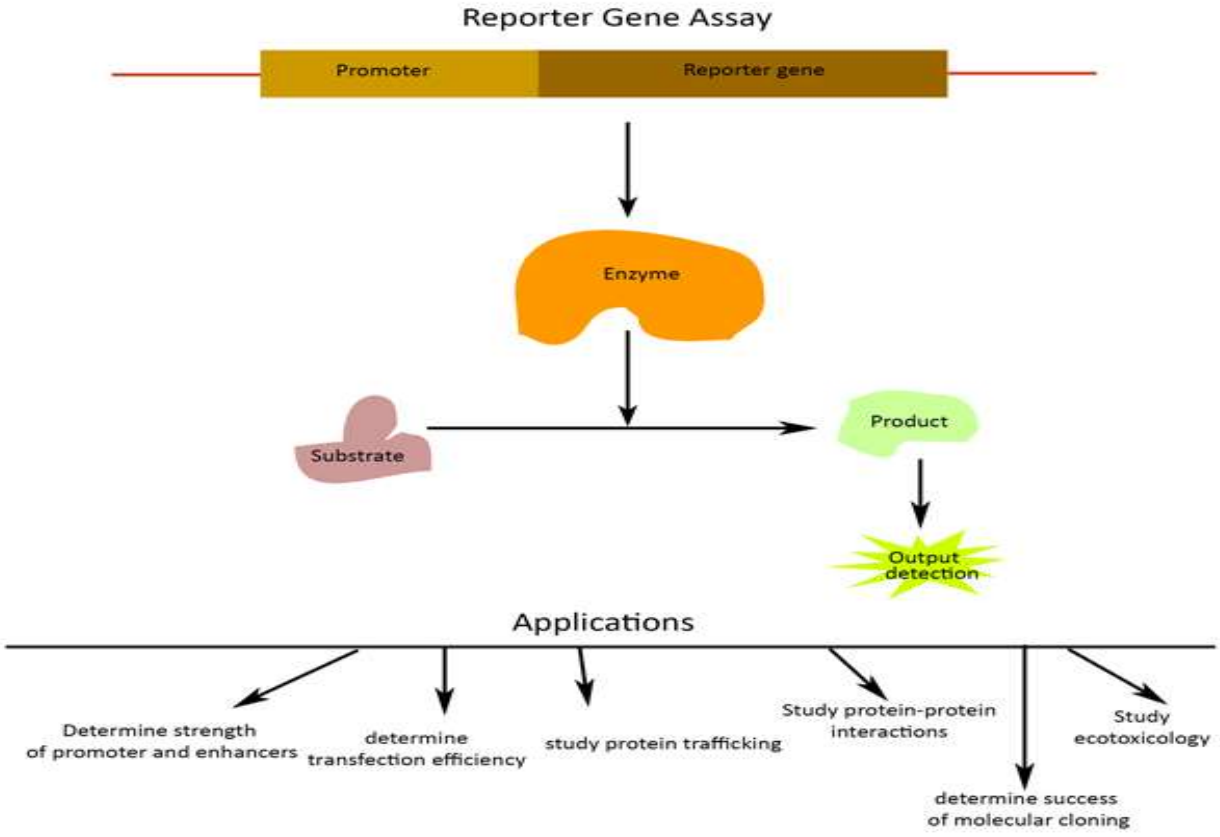
Real-Time PCR

- Steady-state levels of mRNA are quantitated by reverse transcription of the RNA to cDNA followed by quantitative PCR (qPCR) on the cDNA.
- The amount of each specific target is determined by measuring the increase in fluorescence signal from DNA-binding dyes or probes during successive rounds of enzyme-mediated amplification.
- This precise, versatile tool is used to investigate mutations (including insertions, deletions, and single-nucleotide polymorphisms (SNPs)), identify DNA modifications (such as methylation), confirm results from northern blotting or microarrays, and conduct gene expression profiling.

- Expression levels can be measured relative to other genes (relative quantification) or against a standard (absolute quantification). Real-time PCR is the gold standard in nucleic acid quantification because of its accuracy and sensitivity.
- Real-time PCR can be used to quantitate mRNA or miRNA expression following conversion to cDNA or to quantitate genomic DNA directly to investigate transcriptional activity

Promoter Analysis

- Expression of reporter genes/promoter fusions in host cells — promoter activity (transcription rate) is measured in vivo by introducing fusions of various promoter sequences with a gene encoding a product that can be readily measured to monitor activity levels



Most commonly used reporter gene that fits the definition, widely available and commonly used are:

- β-galactosidase [β-Galactosidase Assay (CPRG), Fluorescent β-Galactosidase Assay (MUG)]
- β-glucuronidase (GUS assay used mostly for expression in plants)
- Luciferase (Lumino™ Firefly Luciferase Assay)

- Green fluorescent protein (GFP)
- Secreted Placental Alkaline Phosphatase
- **In vitro transcription (nuclear run-on assays)** — transcription rates are measured by incubating isolated cell nuclei with labeled nucleotides, hybridizing the resultant product to a membrane (slot blot), and then exposing this to film or other imaging media

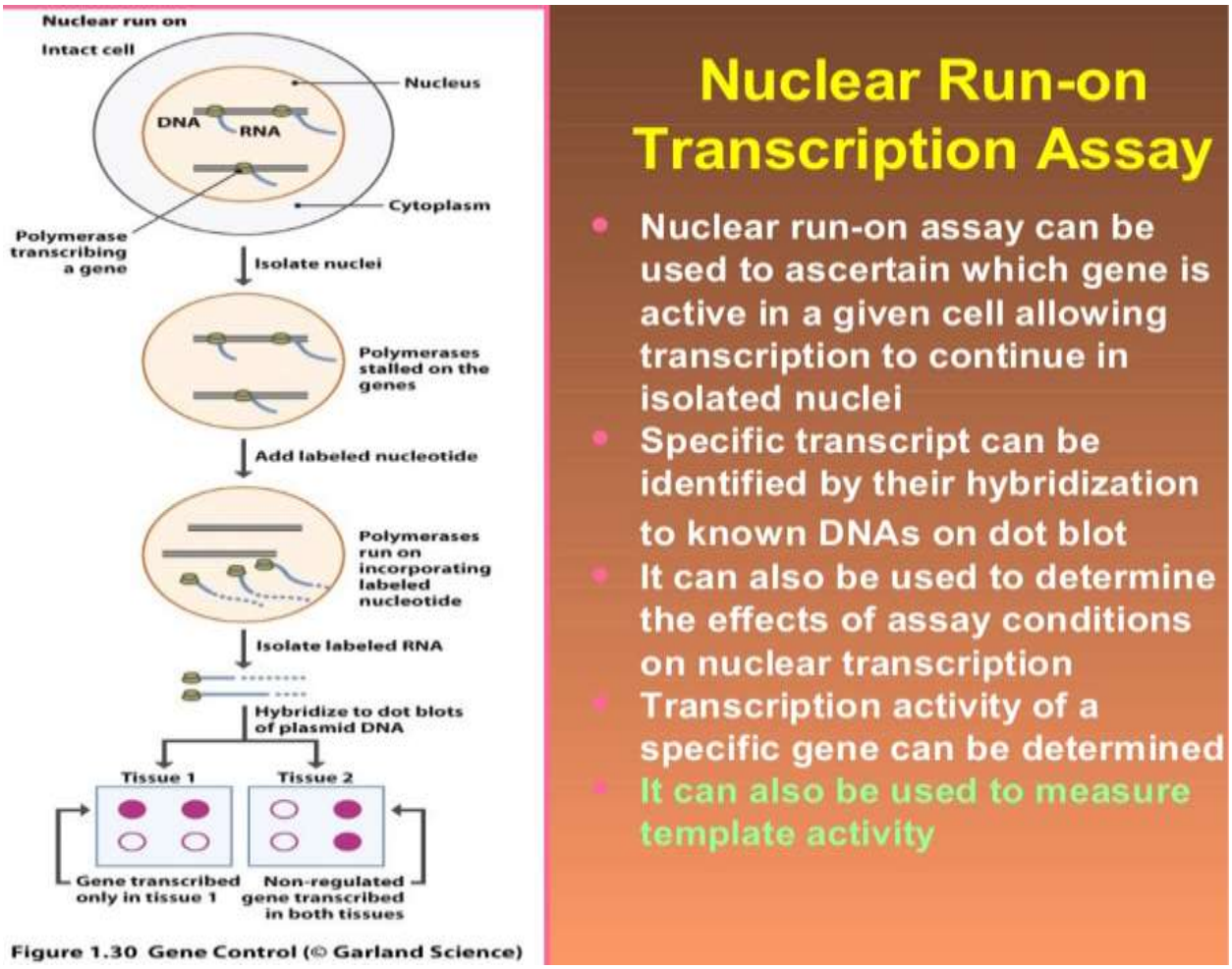
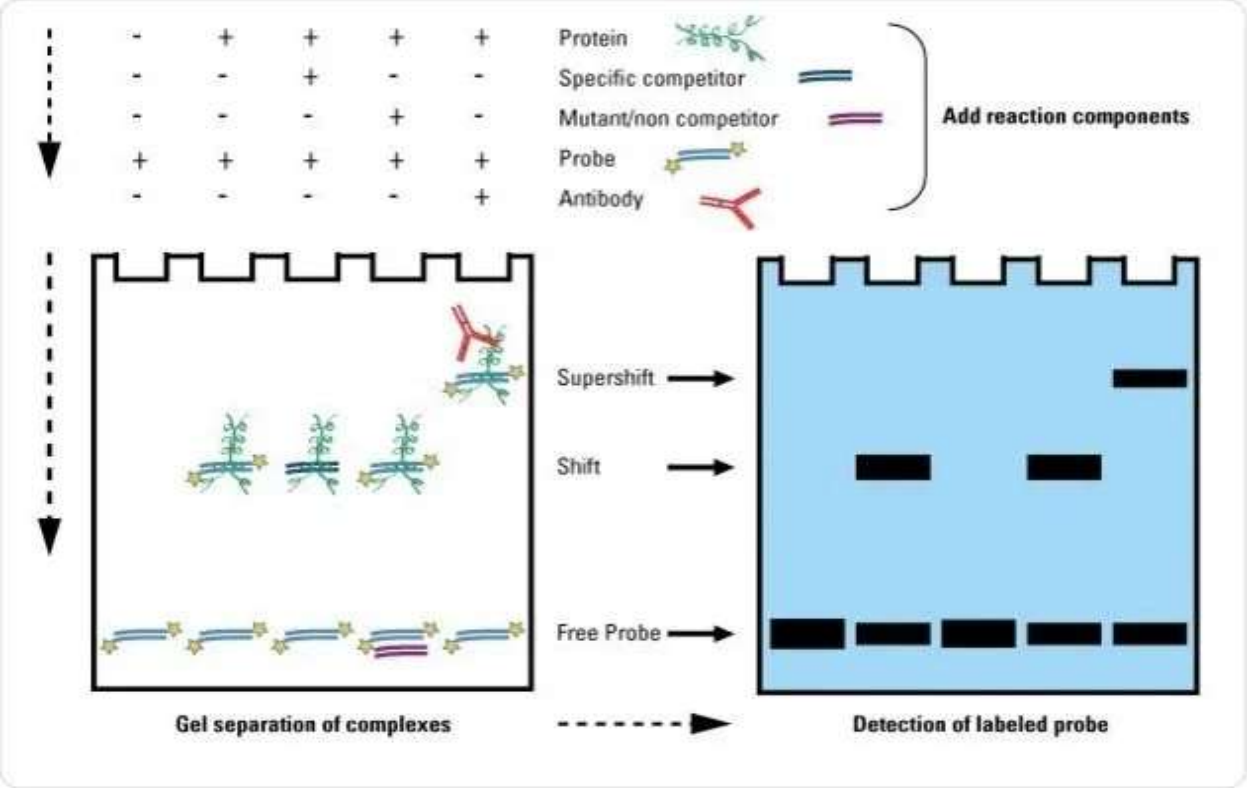


Figure 1.30 Gene Control (© Garland Science)

- **Gel shift assays** — also called electrophoretic mobility shift assays, these are used to study protein-DNA or protein-RNA interactions.
- DNA or RNA fragments that are tightly associated with proteins (such as transcription factors) migrate more slowly in an agarose or polyacrylamide gel (showing a positional shift). Identifying the associated sequences provides insight into gene regulation



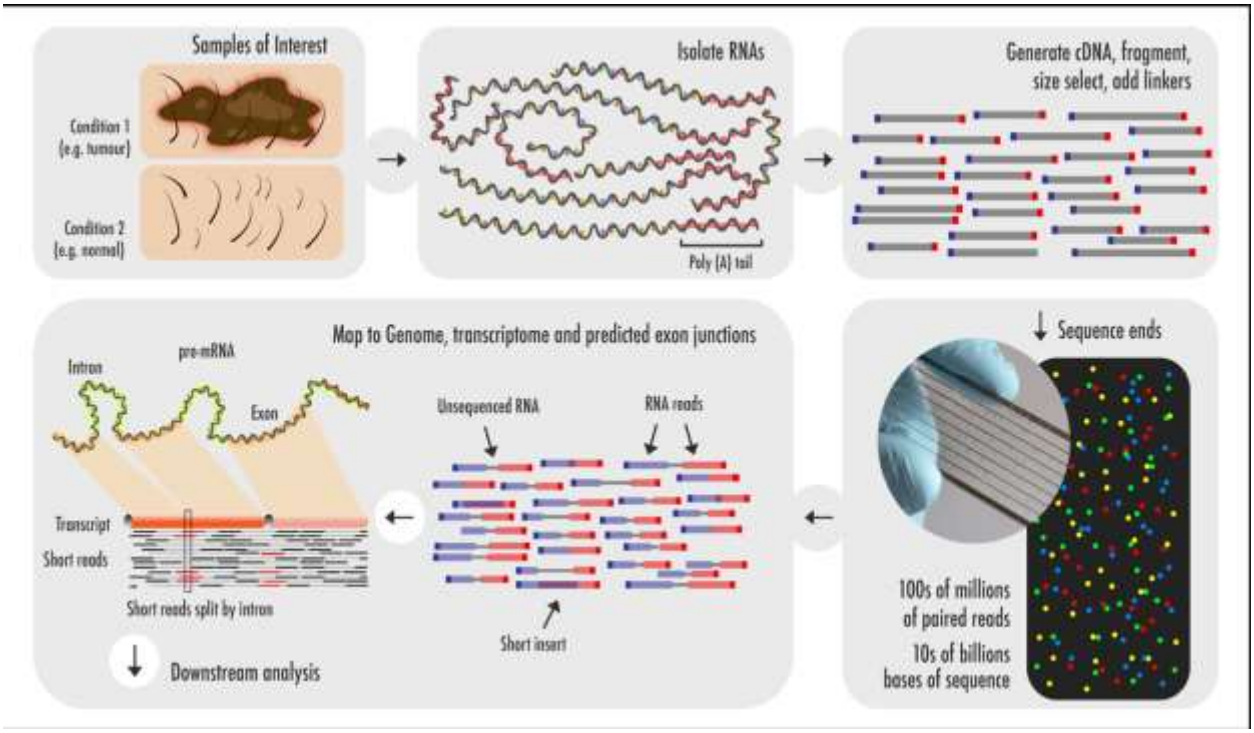
- **Chromatin immunoprecipitation (ChIP) —**
- protein-binding regions of DNA can be identified in vivo. In living cells, DNA and protein are chemically cross-linked, and the resulting complex is precipitated by antibody-coated beads (immunoprecipitation). Following protein digestion and DNA purification, the sequences of the precipitated DNA are determined

Protein Expression

- Western blotting — quantification of relative expression levels for specific proteins is accomplished by electrophoretically separating extracted cell proteins, transferring them to a membrane, and then probing the bound proteins with antibodies (targeted to antigens of interest) that are subsequently detected using various chemistries or radiolabelling
- 2-D Gel Electrophoresis — protein expression profiling is achieved by separating a complex mixture of proteins in two dimensions and then staining to detect differences at the whole-proteome level
- Immunoassays — proteins are quantitated in solution using antibodies that are bound to color-coded beads (as in the Bio-Plex supension array system) or immobilized to a surface (ELISA),

which is subsequently probed with an antibody suspension and is typically detected using a chromogenic or fluorogenic reporter

Posttranslational Modification Analysis

- Immunoassays — levels of protein phosphorylation and other post-translational modifications are detected using antibodies that are specific for these adducts
- Mass spectrometry — proteins and their modifications are identified based on their mass
- **What is RNA-seq?**
- RNA-seq (RNA-sequencing) is a technique that can examine the quantity and sequences of RNA in a sample using next generation sequencing (NGS). It analyzes the transcriptome of gene expression patterns encoded within our RNA. Here, we look at why RNA-seq is useful, how the technique works, and the basic protocol which is commonly used today[1].
- 

What are the applications of RNA-seq?

- RNA-seq lets us investigate and discover the transcriptome, the total cellular content of RNAs including mRNA, rRNA and tRNA. Understanding the transcriptome is key if we are to connect the information on our genome with its functional protein expression.

- RNA-seq can tell us which genes are turned on in a cell, what their level of expression is, and at what times they are activated or shut off. This allows scientists to more deeply understand the biology of a cell and assess changes that may indicate disease.
- Some of the most popular techniques that use RNA-seq are transcriptional profiling, SNP identification, RNA editing and differential gene expression analysis.

  This can give researchers vital information about the function of genes. For example, the transcriptome can highlight all the tissues in which a gene of unknown function is expressed, which might indicate what its role is.
- It also captures information about alternative splicing events (Figure 1), which produce different transcripts from one single gene sequence. These events would not be picked up by DNA sequencing. It can also identify post-transcriptional modifications that occur during mRNA processing such as polyadenylation and 5' capping.

How does RNA-seq work?

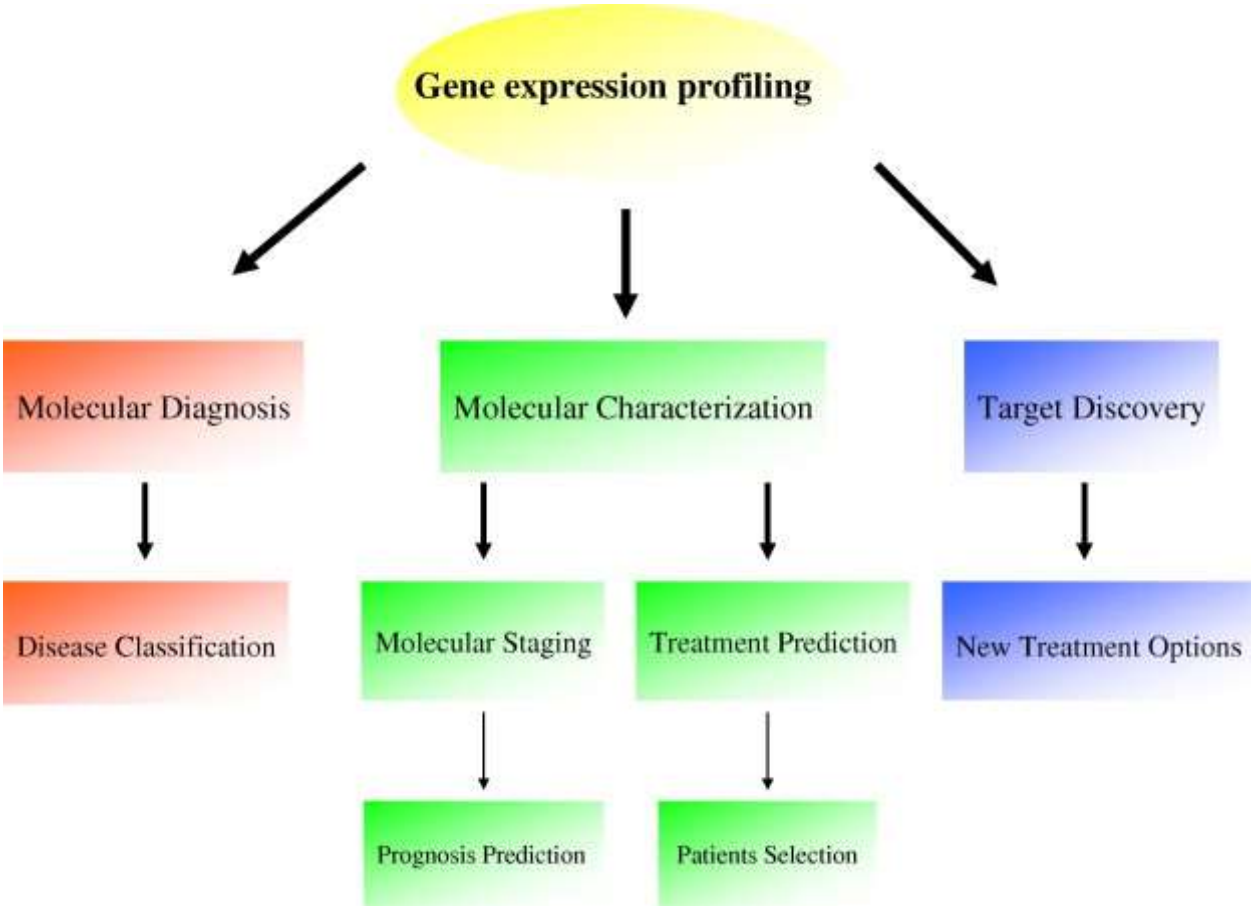
- Early RNA-seq techniques used Sanger sequencing technology, a technique that although innovative at the time, was also low-throughput, costly, and inaccurate. It is only recently, with the advent and proliferation of NGS technology, have we been able to fully take advantage of RNA-seq's potential[4].

  The first step in the technique involves converting the population of RNA to be sequenced into cDNA fragments (a cDNA library). This allows the RNA to be put into an NGS workflow. Adapters are then added to each end of the fragments. These adapters contain functional elements which permit sequencing; for example, the amplification element and the primary sequencing site. The cDNA library is then analyzed by NGS, producing short sequences which correspond to either one or both ends of the fragment. The depth to which the library is sequenced varies depending on techniques which the output data will be used for. The sequencing often follows either single-read or paired-end sequencing methods. Single-read sequencing is a cheaper and faster technique (for reference, about 1% of the cost of Sanger sequencing) that sequences the cDNA from just one end, whilst paired-end methods sequence from both ends, and are therefore more expensive and time-consuming[5,6].

- A further choice must be made between strand-specific and non-strand-specific protocols. The former method means the information about which DNA strand was transcribed is retained. The value of extra information obtained from strand-specific protocols make them the favorable option.

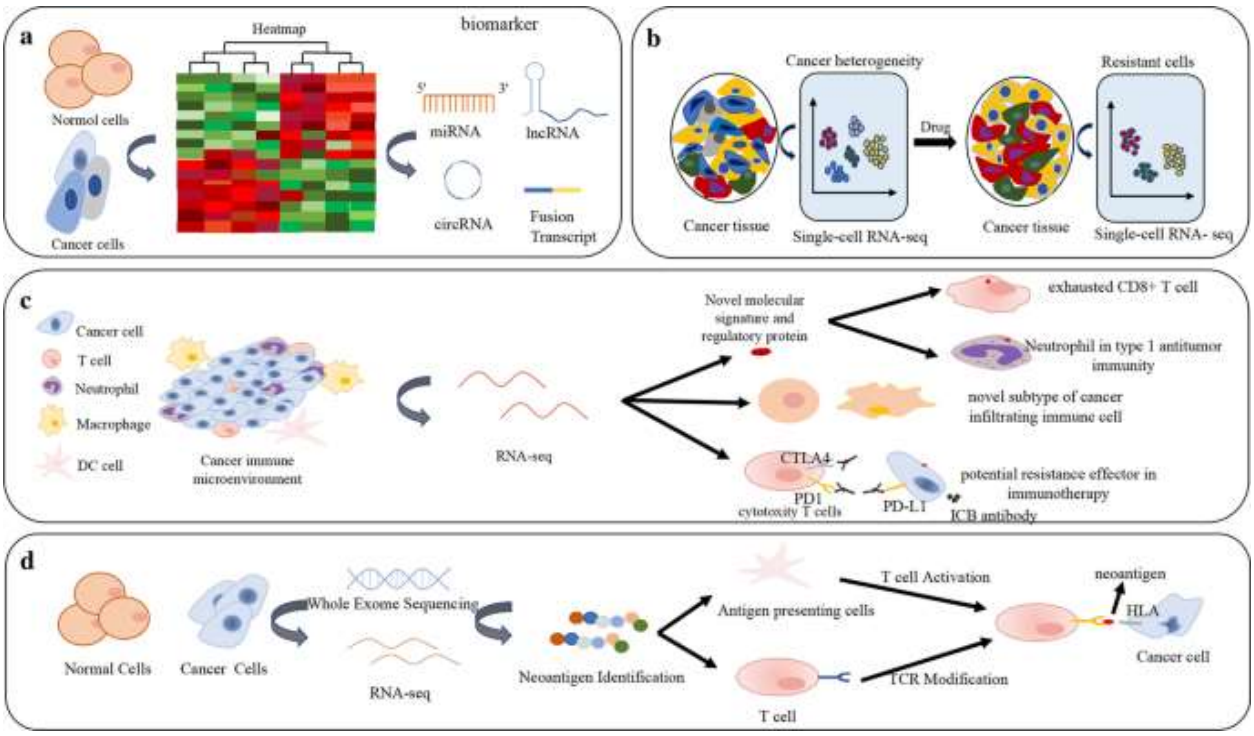
  These reads, of which there will be many millions by the end of the workflow, can then be aligned to a genome of reference and assembled to produce an RNA sequence map that spans the transcriptome[7].

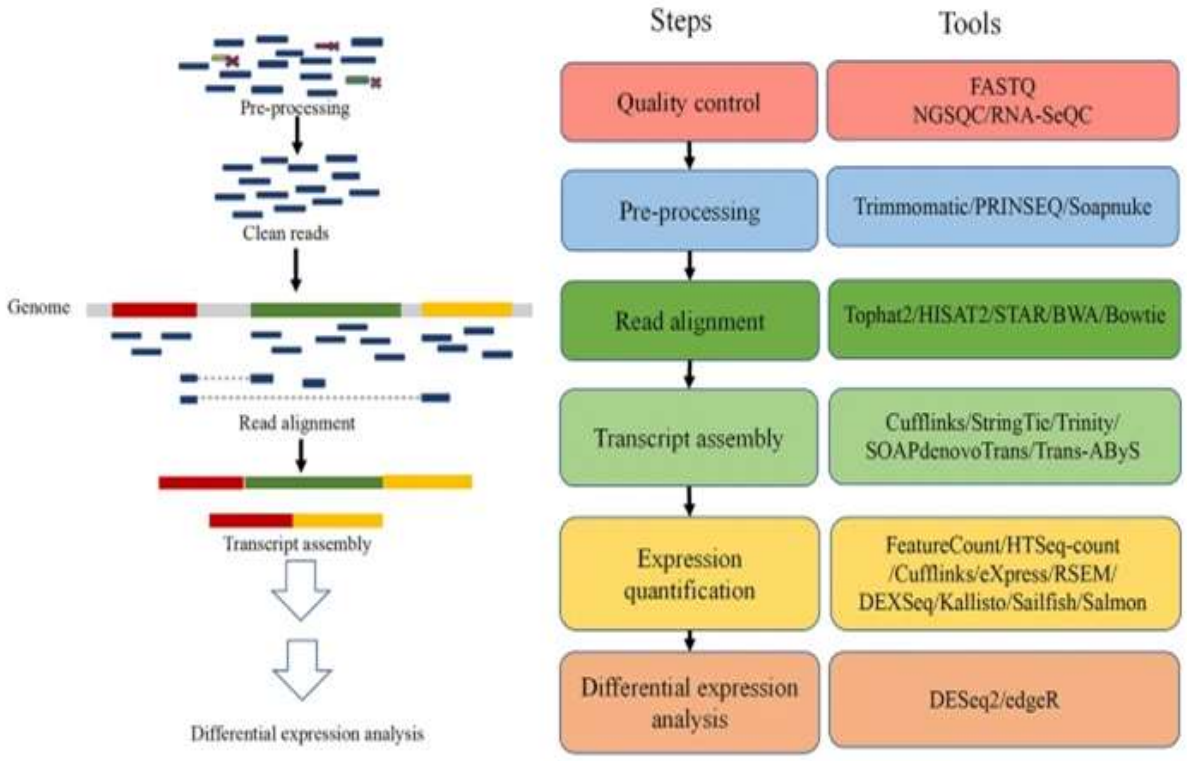
- RNA-seq vs microarrays: Why RNA-seq is considered superior
- RNA-seq is widely regarded as superior to other technologies, such as microarray hybridization. There are several reasons for RNA-seq's well-regarded status
- Not limited to genomic sequences – unlike hybridization-based approaches, which may require species-specific probes, RNA-seq can detect transcripts from organisms with previously undetermined genomic sequences. This makes it fundamentally superior for the detection of novel transcripts, SNPs or other alterations.
- Low background signal – the cDNA sequences used in RNA-seq can be mapped to targeted regions on the genome, which makes it easy to remove experimental noise. Furthermore, issues with cross-hybridization or sub-standard hybridization, which can plague microarray experiments, are not an issue in RNA-seq experiments.
- More quantifiable - Microarray data is only ever displayed as values relative to other signals detected on the array, whilst RNA-seq data is quantifiable. RNA-seq also avoids the issues microarrays have in detecting very high or very low expression levels.
- ***Applications of gene expression studies***



Microarrays – definition , discovery, technique, making microarrays

- Microarray technology is a developing technology used to study the expression of many genes at once.
- It involves placing thousands of gene sequences in known locations on a glass slide called a gene chip.
- A sample containing DNA or RNA is placed in contact with the gene chip. Complementary base pairing between the sample and the gene sequences on the chip produces light that is measured.
- Areas on the chip producing light identify genes that are expressed in the sample.





Bioinformatics tools commonly used in RNA-seq data analysis. These tools are primarily used in the four main processes of RNA-seq data analysis, including quality control, read alignment and transcript assembly, expression quantification and differential expression analysis

Examples of gene expression in therapeutic development

Rapidly accelerating the identification of candidate therapeutic targets

- Genentech identified candidate therapeutic targets for invasive prostate cancer (PNAS, March 2002)[3]

Improving the ability to prioritise potential therapeutic targets

- Stanford validates new targets in autoimmune encephalomyelitis (Nature Medicine, May 2002)[27]

Understanding mechanism of drug action

- UCB Pharma determined mechanism of action of effective anti-epileptic drug, levetiracetam (European Journal of Neuroscience, 2004)[2]

Predicting the human toxicity of novel compounds

- Identification of gene profiles that predict human specific toxicity using rat gene expression data Toxicology Science, 2002)[31]

Understanding the mechanism of toxicity of compounds

- Mechanism of canine-specific hepatic sclerosis determined for novel compound (Mattes, Gene Logic, submitted for publication)

Identifying biomarkers used to assess therapeutic response

- Blood biomarkers identified to follow response to anti-depressants in treating patients with major depression syndrome (American Journal of Medical Genetics, 2005[1] and unpublished communication)

Improving clinical trial outcomes by selecting appropriate patients for investigational new drugs

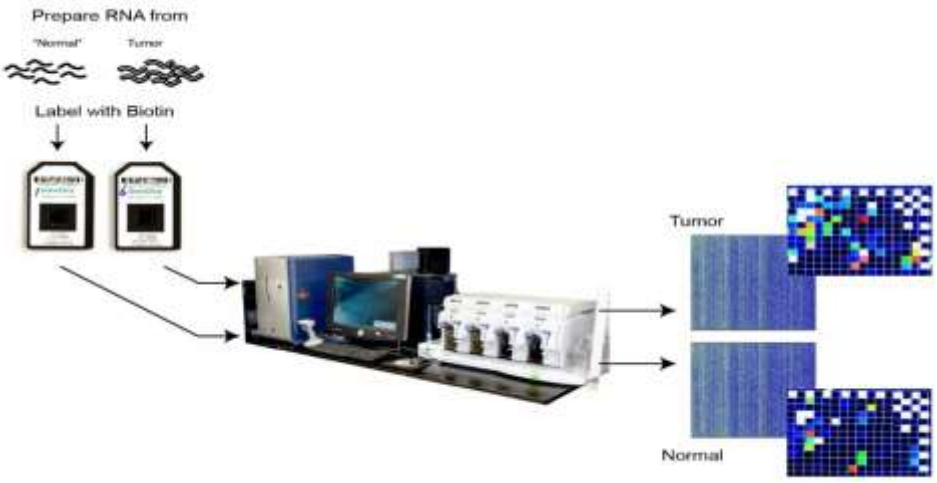
- Clinical stratification of tomoxifen-treated, node negative breast cancer (NEJM, December 2004)[13]

Improving disease classification for developing targeted therapeutics

- Diffuse large B-cell lymphoma redefined based on expressions profile (Nature Medicine, January 2002)[25]

Predicting biological response to novel chemical structures

- NIH correlates molecular chemical substructure with biological response predicted from gene expression data (The Pharmacogenomics Journal, 2002)[28]



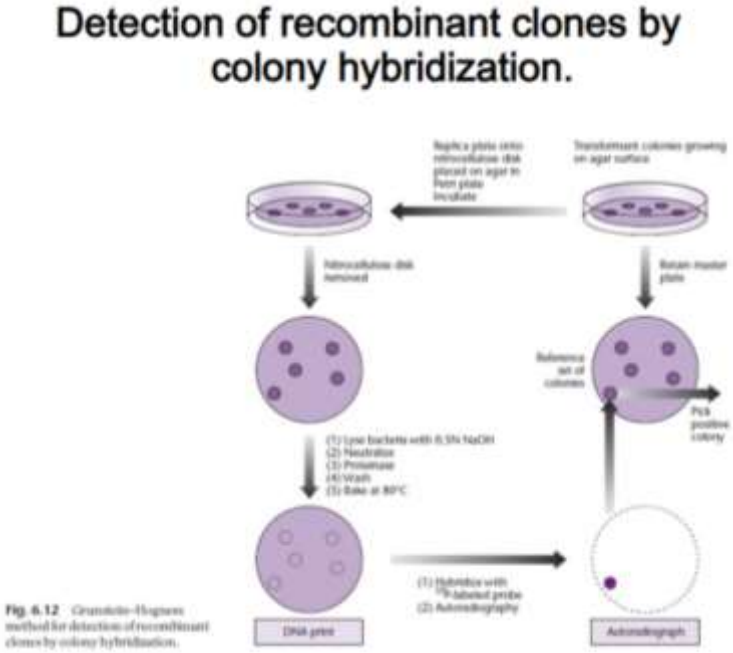
Microarrays – definition , discovery

Early DNA arrays

- After the first description of the double helix DNA structure by Watson and Crick in 1953, the process of separating the two strands was soon reversed with methods of DNA molecular hybridization quickly explored.
- Molecular hybridization is the occurrence of single-stranded DNA binding to complimentary DNA. The complimentary base pairs that form the structure of the opposite strands of DNA are the foundation for all analysis methods involving DNA sequences.

- In 1975, Grunstein and Hogness applied the process of molecular hybridization to DNA released from blotted microbial colonies, a useful process for screening bacteria clones.

The colony hybridization method was formed by randomly cloning *Escherichia coli* (*E. coli*) DNA onto agar petri plates covered with nitrocellulose filters. A radioactively labeled probe was then added which would bind to complimentary DNA within the sample.



The aforementioned method formed a random orientation of sample DNA spots representing the cloned fragments of DNA. This is an early example of a labeled probe being utilized in order to identify complimentary base pair binding. It can therefore be considered as one of the first examples of a DNA array.

- *Gergan et al. adapted this methodology* to produce arrays in 1979. Multiple plates on agar were replicated to produce arrays through the use of a mechanical 144 pin device for placing samples in the corresponding amount of well microplates. This allowed for the production of arrays for over a thousand different bacterial colonies.

- The colonies could then be easily transferred to paper filters for the necessary lysis, denaturation and fixing steps for producing hybridized DNA. The technology of filter based arrays was used in research that led to the identification of single nucleotide polymorphisms (SNPs) and the ability to clone specific genes of interest.
- The ability to analyze multiple hybridization targets was automated in the late 1980s and early 1990s. Robotic technology was used to quickly array clones from microtiter plates onto filters. The arrays created a defined pattern allowing parallel hybridizations to be produced.
- Efficiency was increased with the errors that occur during repetitive procedures being reduced through the automated placing of samples on the array. The increased speed and accuracy from automation was an important step in the development towards microarrays.
- A further development of complimentary DNA (cDNA) cloning was also an important foundation for the microarray, as it led to the creation of reference sets of cDNA and corresponding filter arrays for whole genomes.
- In 1995, the first study that used the word 'microarray' was published which explained how the expression of many genes could be monitored in parallel through the use of this new technology. The sample array was constructed through high-speed robotic printing of cDNA on glass.
- The small size of spots on the array and high density of the arrays produced hybridization volumes of two microliters, which was the volume that enabled the detection of rare transcripts within the probe samples.
- The microarray was a technical advancement that meant a broader examination of gene expression could be accomplished. In 1997, the researchers from Stanford University published the first whole-genome microarray study of gene expression by placing the whole yeast genome on a microarray.
- **History**

  Earliest form of microarray is the Southern blot, developed in 1975 by Dr. Edward Southern of Edinburgh University
- In this technique, fragmented DNA is bound to a substrate (often a nitrocellulose or nylon membrane), denatured, dried and then exposed to a labeled hybridization probe in an appropriate buffer
- Blot is then extensively washed and analyzed by Xray film, autoradiography or membrane chromogen detection, depending on the type of probe label employed

- Southern blotting has been largely replaced by newer molecular techniques but it has value in analyzing several trinucleotides repeat syndromes (Fragile X syndrome, Huntington chorea), where the length of the expanded DNA is greater than the usual amplification ability of PCR
- Array technology was used by Augenlicht et al. in 1984 to analyze retroviral long terminal repeat (LTR element expression in murine colon tumors (J Biol Chem 1984; 259:1842)
- In 1987, Kulesh et al. used arrays to analyze the expression of more than 2,000 different genes constructed from a human fibrosarcoma cell line, with and without interferon treatment (Proc Natl Acad Sci USA 1987;84:8453)
    - Different mRNA derived cDNAs were spotted onto filter paper and analyzed
    - 29 sequences were induced by interferon treatment
- **Miniaturized microarrays were introduced in 1995 (Science 1995; 270:467)**
- First complete eukaryotic genome was placed on microarray in 1997, when Lashkari et al. placed a maximum of 2,470 open reading frames on a glass slide and analyzed total mRNA expression (cDNA) in *S. cerevisia*, examining the effects of heat and cold shock and culture in glucose vs galactose on global gene expression profiles (Proc Natl Acad Sci USA 1997;94:13057)
- Since its first research use in the 1980s, the development of better surface technologies, more powerful robots for arraying, better nucleic acid dye labeling techniques and improved computational power and automated analyzers have vastly improved the power and efficiency of microarray, while also lowering the cost of these analyses
- Microarray is currently used to analyze many different systems, including the classification of microbes and human microbial pathogens, cellular responses to pathogens, drug and toxic exposures, tumor classification, single nucleotide polymorphism detection, the detection of gene fusions, comparative genomic hybridization, alternative splicing detection (exon junction array / exon arrays) and gene expression profiling via analyzing global mRNA levels
- Most microarray protocols use reverse transcriptase to convert mRNA into cDNA, as DNA is more stable with RNA
- DNA microarrays, also called DNA chips, gene chips, DNA arrays, gene arrays and biochips, are microscopic slides of glass or silicon printed with thousands of small spots in grid fashion with each containing known DNA or gene.

- Each slide acts as probe to detect gene expression.
- Basically, it ‘s been evolved from southern blotting .It is different from Southern blotting as here the probe is fixed/attached and sample DNA is labeled rather than probe.

PRINCIPLE

- DNA mICROARRAY
- The basic principle behind DNA microarray lies on Nucleic acid hybridization.
- During this method, two complementary strands of DNA are joined together by hydrogen bond to make a double stranded molecule by hydrogen bond.
- Restriction endonuclease is employed to cut the unknown DNA molecules into small fragments. Fluorescent markers are attached to the fragments and these get react with probes in DNA chips.
- DNA probes are then binds with the target DNA with complementary sequences and unbounded DNA fragments are washed away.
- Identification of the target pieces of DNA is done by their fluorescene emission passing through a laser beam and computer recorded the pattern of emission as well as DNA identification.

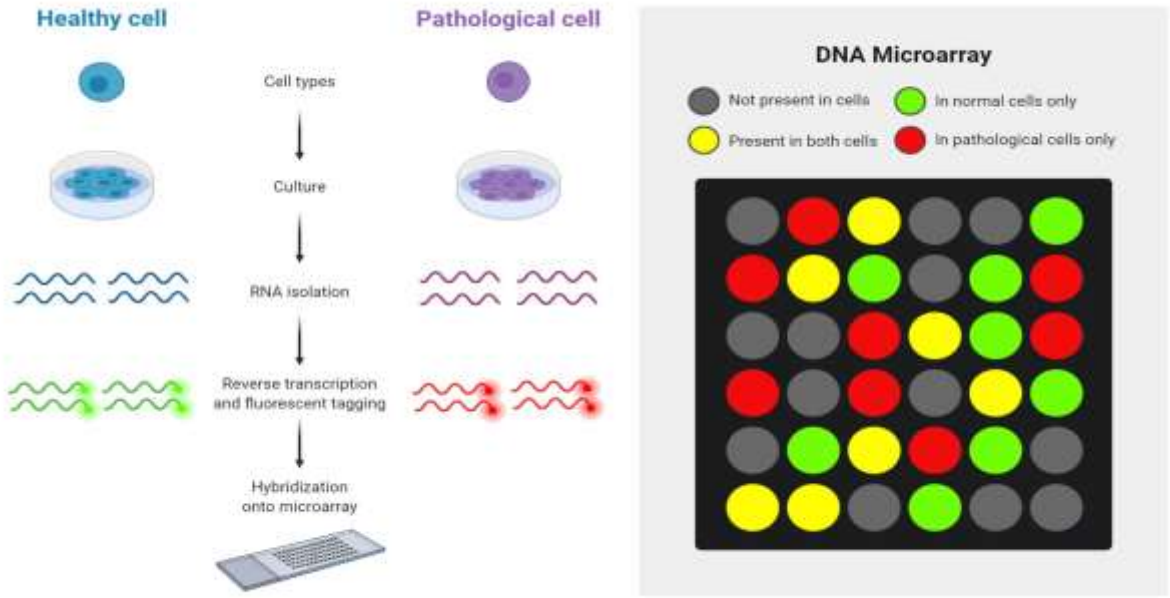


Image By Sagar Aryal, created using biorender.com

TYPES OF DNA MICROARRAY

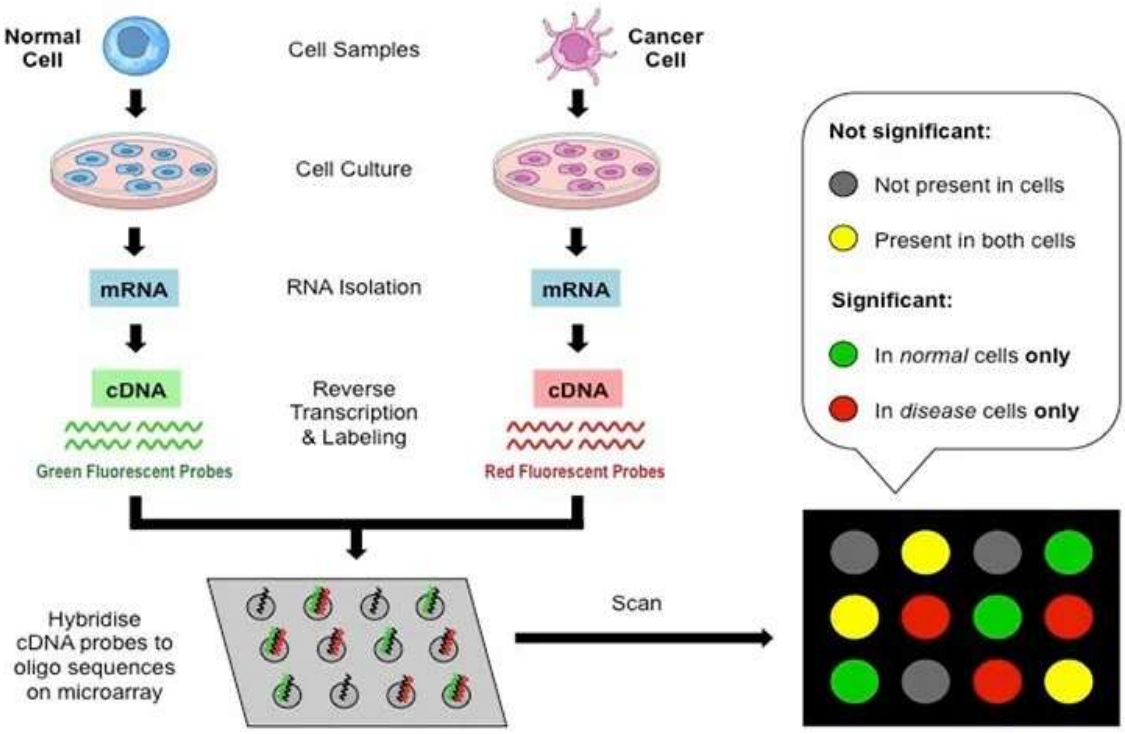
There are four types of DNA microarray:

1. Oligo DNA microarray: It uses oligonucleotides of 20-50 nucleotides long. Oligonucleotides are synthesized directly on the slide. Single color hybridization used for each probe. It has good specifcity but poor sensitivity.
2. cDNA microarray: It is usually referred to as spotted microarray in within which DNA fragments of any length (500bp-1kb) or oligos of 20-100 nts are stuck to the glassslides. It uses two colors hybridization for every probe.
3. BAC Microarray: It uses the template which is amplified by polymerase chain reaction as the probe.
4. SNA Microarray: It is used to detect polymorphisms within a population.

REQUIREMENTS:

1. DNA Chip
2. Target sample
3. Sample
4. Enzymes
5. Fluorescent dyes
6. Probes

- DNA microarrays are solid supports, usually of glass or silicon, upon which DNA is attached in an organized pre-determined grid fashion.
- Each spot of DNA, called a probe, represents a single gene.
- DNA microarrays can analyze the expression of tens of thousands of genes simultaneously.
- There are several synonyms of DNA microarrays such as DNA chips, gene chips, DNA arrays, gene arrays, and biochips.



Principle of DNA Microarray Technique

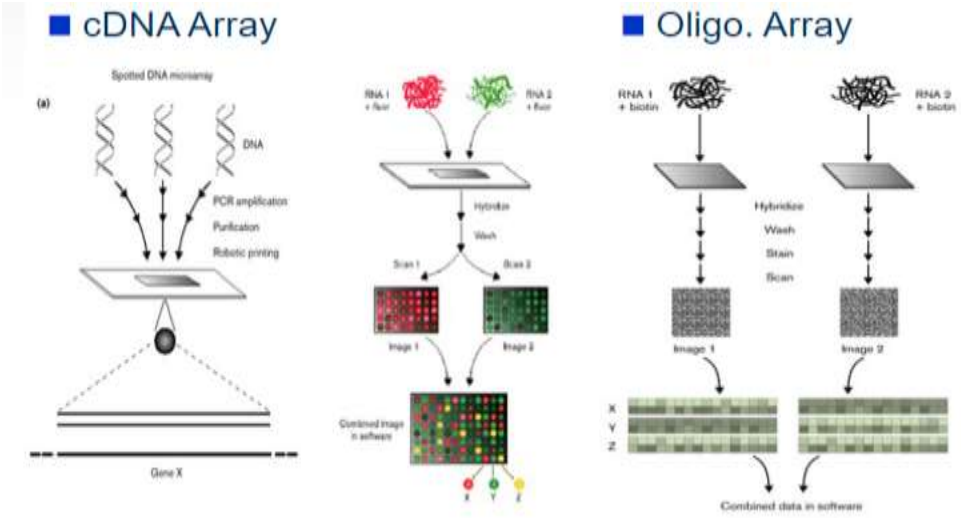
- The principle of DNA microarrays lies on the hybridization between the nucleic acid strands.
- The property of complementary nucleic acid sequences is to specifically pair with each other by forming hydrogen bonds between complementary nucleotide base pairs.
- For this, samples are labeled using fluorescent dyes.
- At least two samples are hybridized to chip.
- Complementary nucleic acid sequences between the sample and the probe attached on the chip get paired via hydrogen bonds.
- The non-specific bonding sequences while remain unattached and washed out during the washing step of the process.
- Fluorescently labeled target sequences that bind to a probe sequence generate a signal.
- The signal depends on the hybridization conditions (ex: temperature), washing after hybridization etc while the total strength of the signal, depends upon the amount of target sample present.
- Using this technology, the presence of one genomic or cDNA sequence in 1,00,000 or more sequences can be screened in a single hybridization.

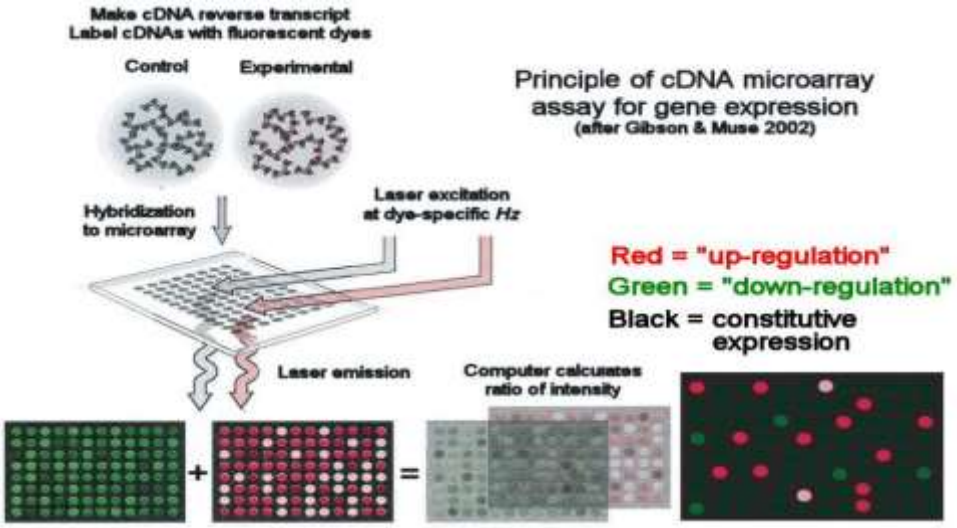
There are 2 types of DNA Chips/Microarrays:

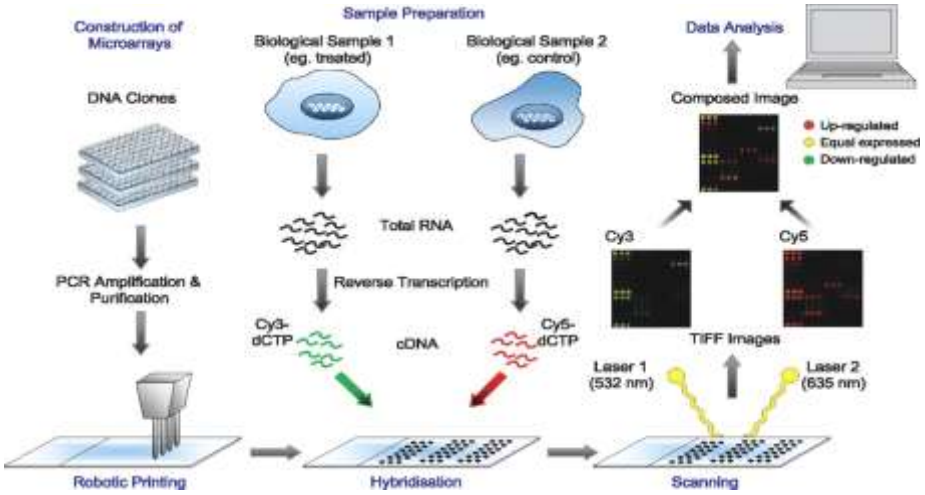
1. cDNA based microarray
2. Oligonucleotide based microarray

Spotted DNA arrays ("cDNA arrays")

- Chips are prepared by using cDNA.
- Called cDNA chips or cDNA microarray or probe DNA.
- The cDNAs are amplified by using PCR.
- Then these immobilized on a solid support made up of nylon filtre of glass slide (1 x 3 inches). The probe DNA are loaded into a a spotting spin by capillary action.
- Small volume of this DNA preparation is spotted on solid surface making physical contact between these two.
- DNA is delivered mechanically or in a robotic manner.


cDNA Array
Oligo. Array
Spotted DNA microarray
DNA
PCR amplification
Purification
Robotic printing
Gene X
RNA 1
+ fluor
RNA 2
+ fluor
Hybridize
Wash
Scan 1
Scan 2
Image 1
Image 2
Combined image in software
RNA 1
+ biotin
RNA 2
+ biotin
Hybridize
Wash
Stain
Scan
Image 1
Image 2
X
Y
Z
Combined data in software


Make cDNA reverse transcript
Label cDNAs with fluorescent dyes
Control
Experimental
Hybridization to microarray
Laser excitation at dye-specific Hz
Laser emission
Computer calcuates ratio of intensity
Principle of cDNA microarray assay for gene expression
(after Gibson & Muse 2002)
Red = "up-regulation"
Green = "down-regulation"
Black = constitutive expression


Construction of Microarrays
DNA Clones
PCR Amplification & Purification
Robotic Printing
Sample Preparation
Biological Sample 1 (eg. treated)
Biological Sample 2 (eg. control)
Total RNA
Reverse Transcription
Cy3-dCTP
cDNA
Cy5-dCTP
Hybridisation
Data Analysis
Composed Image
Up-regulated
Equal expressed
Down-regulated
Cy3
Cy5
TIFF Images
Laser 1 (532 nm)
Laser 2 (635 nm)
Scanning

Oligonucleotide arrays (Gene Chips)

- In oligonucleotide microarrays, short DNA oligonucleotides are spotted onto the array.
- Small number of 20-25mers/gene.
- The main feature of oligonucleotide microarray is that each gene is normally represented by more than one probe.
- Enabled by photolithography from the computer industry
- Off the shelf

Some key issues involved in microarray analysis

Parameter Issue

- **Experimental design**

- ❑ Consider the biological question(s) and the ability to achieve statistical significance
- ❑ Seek expert statistical advice during the early planning stages
- ❑ Microarray experiments have multiple sources of variation and must be carefully controlled
- ❑ Biological and technical replication are essential
- ❑ Sample pooling should be avoided if accurate sample synchronisation is not possible
- ❑ Microarray analysis of purified cells will only reveal genes expressed by these cells, but removal from the in vivo microenvironment may alter gene expression
- ❑ There are limitations in the use of both whole tissue and purified cells, which may necessitate the use of microdissection and RNA amplification techniques
- ❑ When using clinical samples, detailed patient history and tissue histopathology are critical to the interpretation of gene expression profiles

Target RNA preparation

- ❑ The quality of the target RNA is one of the most important factors in the success or failure of a microarray experiment

Data analysis

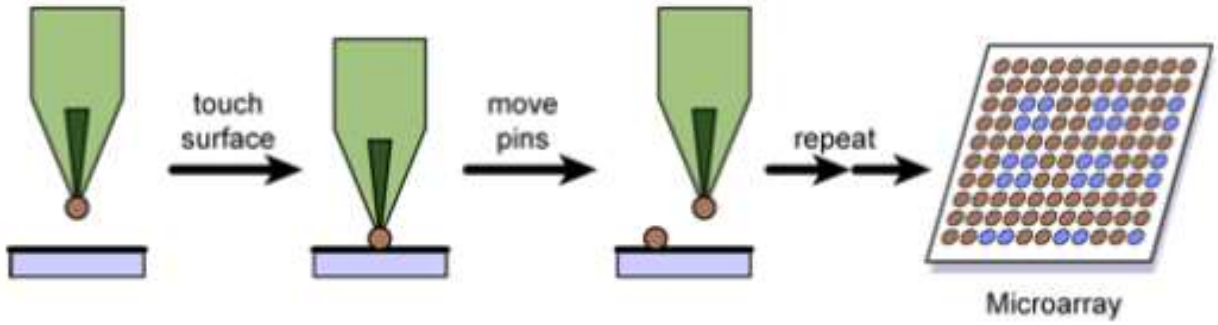
- ❑ While critical to the outcome of a microarray experiment, statistical analysis of microarray data is not well understood by many biologists and expert advice should be sought

Data validation

- ❑ The biomedical research community does not yet accept that microarray data can stand alone without independent validation
- ❑ The investigator must decide which genes to examine further, and those with larger fold changes and statistical significance are often the best candidates
- ❑ To describe a biological event or system, gene expression data obtained by microarray analysis must be extended to the study of protein products

Spotted DNA Microarrays

- The first DNA microarrays were spotted with probes that were made by oligonucleotide chemical synthesis and then attached to the array.
- These probes have to be "spotted" or "printed" using a robot onto a very fine grid by a sort of specialized inkjet-like printer, which uses the same technology as computer printers to expel nanoliter to picoliter volume droplets of probe solution, instead of ink, onto the slide.
- Alternatively, these probes can be applied with a pin directly onto a specific location on the surface. The number of spots (aka features) applied onto the DNA microarray is limited to prevent cross-contamination problems.

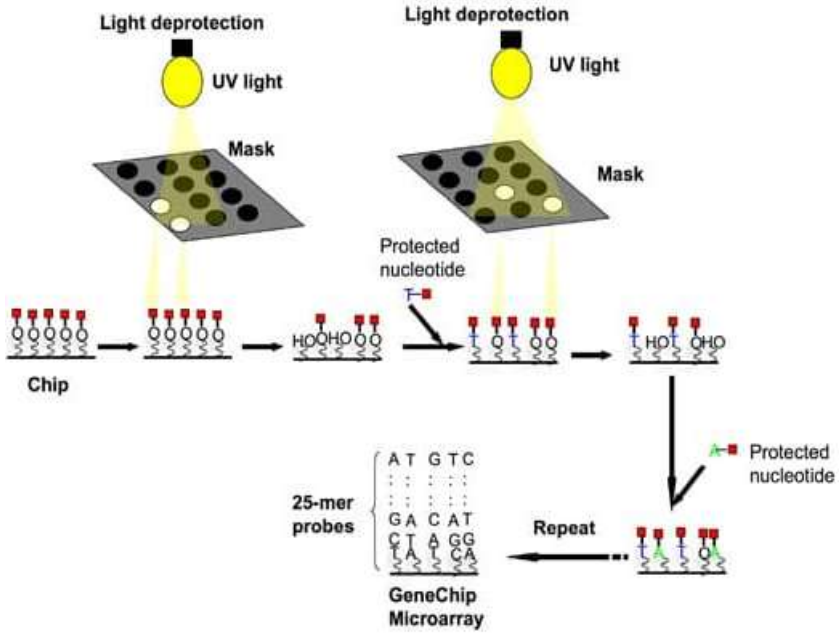


- Affymetrix platform include Agilent arrays that use an inkjet spotting process for in-situ oligonucleotide synthesis, using five "ink" printing of the 4 nucleotide precursors plus catalyst, combined with coupling and deprotection steps that do not require use of photolithographic masks.

This technology relies on printing picoliter volumes of nucleotides on the array surface in repeated rounds of base-by-base printing that extends the length of specific oligonucleotide probes.

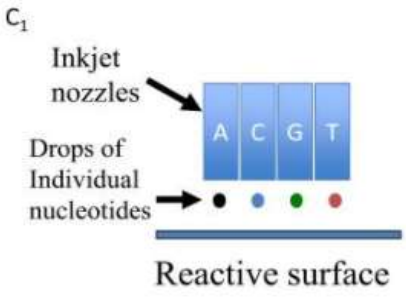
This approach therefore allows synthesis of longer molecules (60-mer length oligos) for their probes. Increased length improves specificity of probes but at increased complexity of design, which reduces the number of features (Affymetrix chips usually feature $>10^6$ spots per microarray, compared to 0.24×10^6features for Agilent). In Situ-Synthesized DNA Microarrays

- In situ-synthesized arrays are high-density oligonucleotide probe DNA microarrays, with Affymetrix GeneChip arrays being the most common.
- These are made using photolithography, which literally means to use light to create a pattern.
- The method relies on UV masking and light-directed combinatorial chemical synthesis on a solid support to selectively synthesize probes directly on the surface of the array, one nucleotide at a time per spot, for many spots simultaneously.
- This process works in the following way: a solid support contains covalent linker molecules that have a protecting group on the free end that can be removed by light.



Affymetrix's proprietary photolithography process for creating DNA microarrays uses a series of photolithographic masks, light deprotection reactions and nucleotide coupling. During each deprotection step, a specific mask is used with particular transparent "windows" to allow the light from a single UV source to deprotect spots, or features on the array to receive a nucleotide.

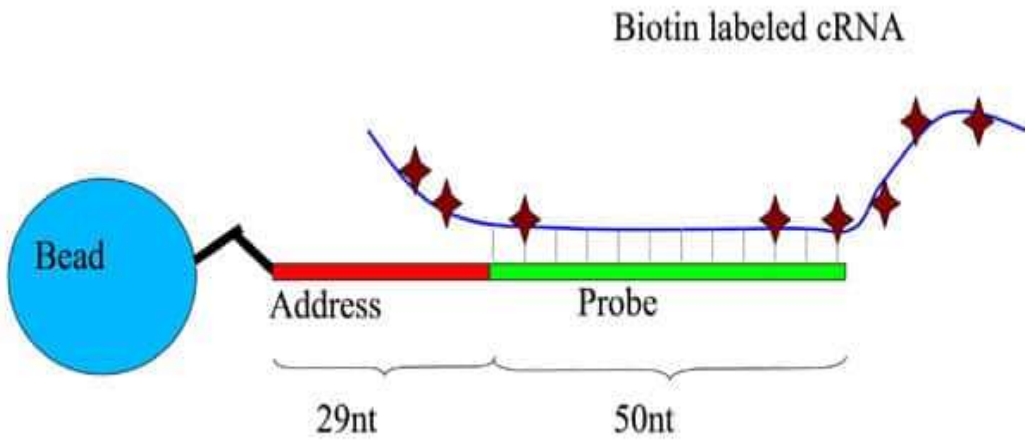
- UV light is directed through a photolithographic mask to deprotect and activate selected sites with hydroxyl groups that initiate coupling with incoming protected nucleotides that attach to the activated sites.
- The mask is designed in such a way that you can choose the exposure sites, and thus specify the coordinates on the array where each nucleotide will be attached. The process is repeated, a new mask is applied activating different sets of sites and coupling different bases, allowing arbitrary DNA probes to be constructed at each site. This process is used to synthesize hundreds of thousands of different oligonucleotides.
- However, it is the length of oligos, not their number, that determines the number of steps required, since many different sites could be synthesized simultaneously.
- Each probe on the chip requires four masks per round of synthesis: one mask to allow addition of the required base and three other masks to prevent light from deprotecting the same spot while the other three nucleotides are being added. On average, each probe is 25 nucleotides long, requiring about 100 masks per chip!!! These microarrays generally employ multiple probes for each gene to improve specificity and feature a match/mismatch probe pair that enable the discrimination of single mismatched base pairs.



High-Density Bead DNA microarrays

- Another type of high-density DNA microarray are the BeadArrays manufactured by Illumina.
- Illumina's Bead Array Technology is based on color-coded 3-micron silica beads that randomly self assemble in either a fiber-optic bundle substrate that then themselves assemble into arrays, or a silica slide substrate.

- When randomly assembled on one of these two substrates, the beads have a uniform spacing of approximately 5.7 microns, with a packing density of about 40,000 array elements per square millimeter.
- This gives the Bead Array platform about 400 times the information density of a typical spotted array.
- Each bead is covered with hundreds of thousands of copies of a specific oligonucleotide that act as the capture sequences in one of Illumina's assays. Each bead has a 23-mer oligo "address" attached to it, which then anchors a 50-mer sequence-specific oligo probe.



- The beads are randomly scattered across etched substrates during the array production process, with each array bundle containing about 50,000 beads.
- With this platform design, a specific oligonucleotide sequence is assigned to each bead type, but is replicated about 30 times on the array at random positions.
- Each gene is represented by two probe sequences. A series of decoding hybridizations are used to determine which oligos are present at each matrix coordinate for every array.

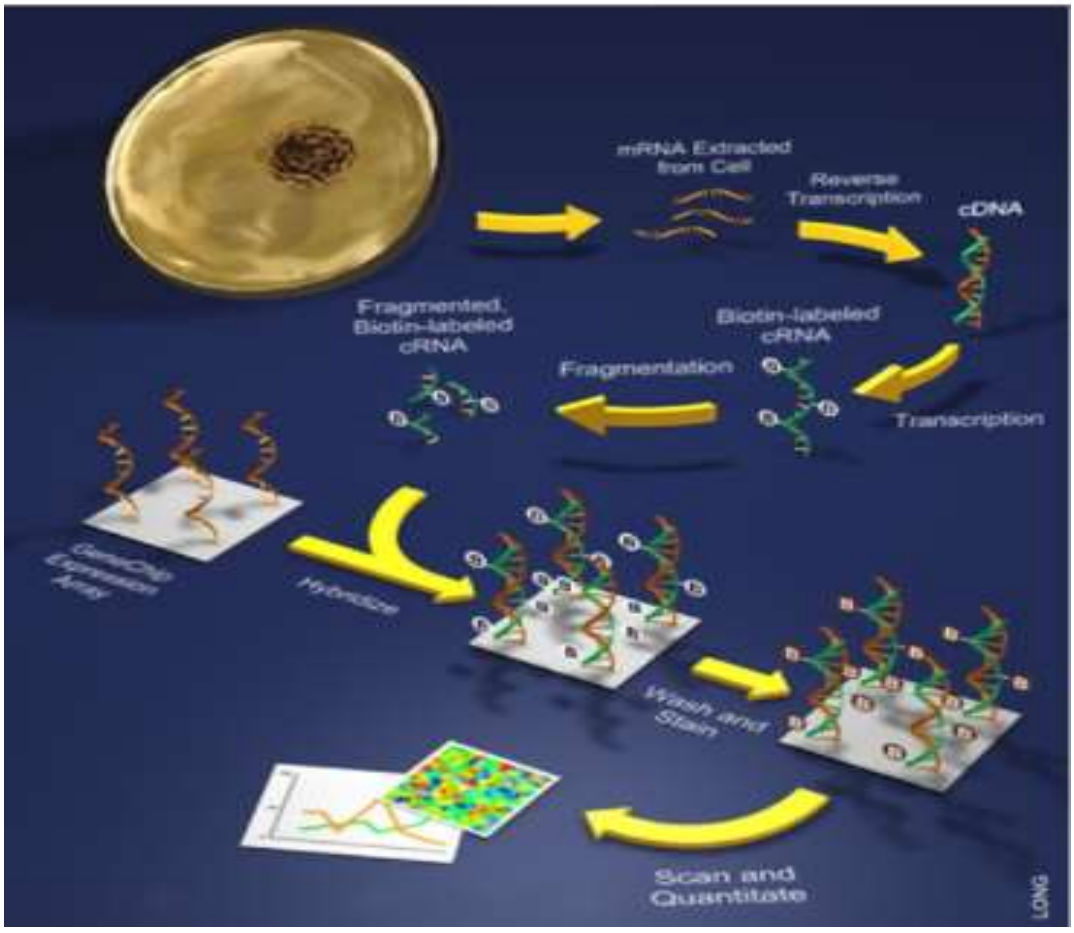
Why use microarrays?

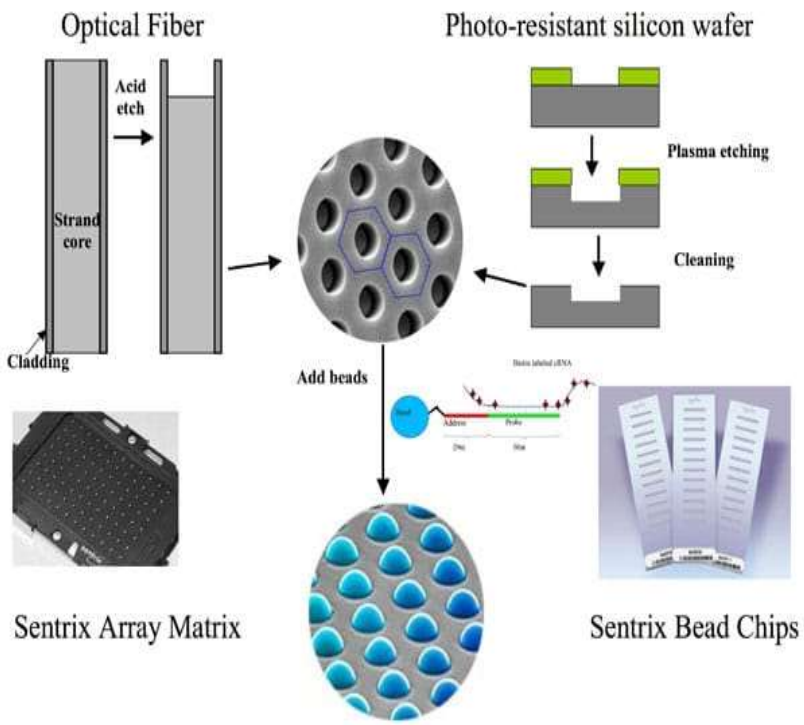
Permits expression profiling of thousands of genes in parallel

Why use Affymetrix microarrays?

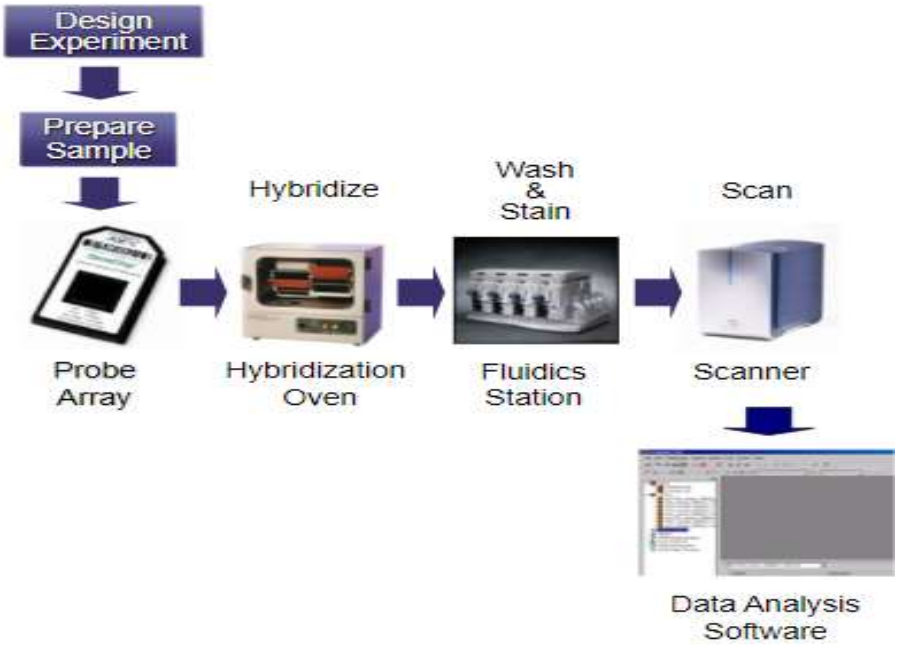
- Well established platform
- Arrays for various species
- Enables combination of different applications
- Support from the bioinformatic community
- Well annoted probesets

Steps in analysis

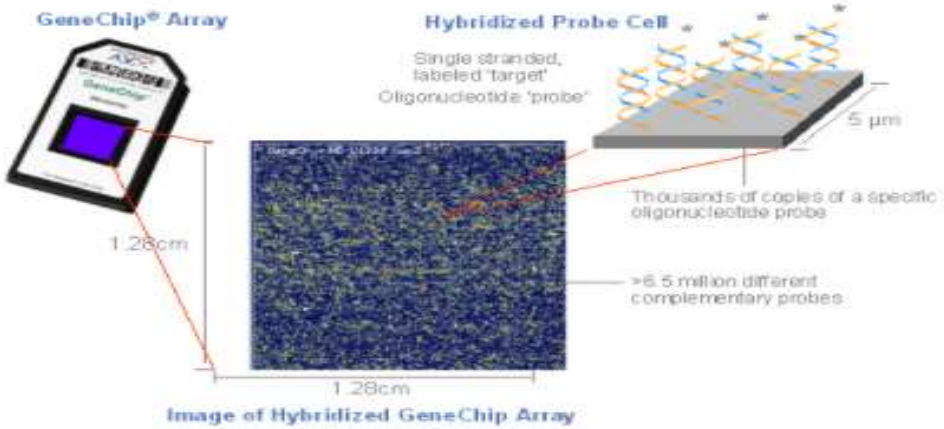




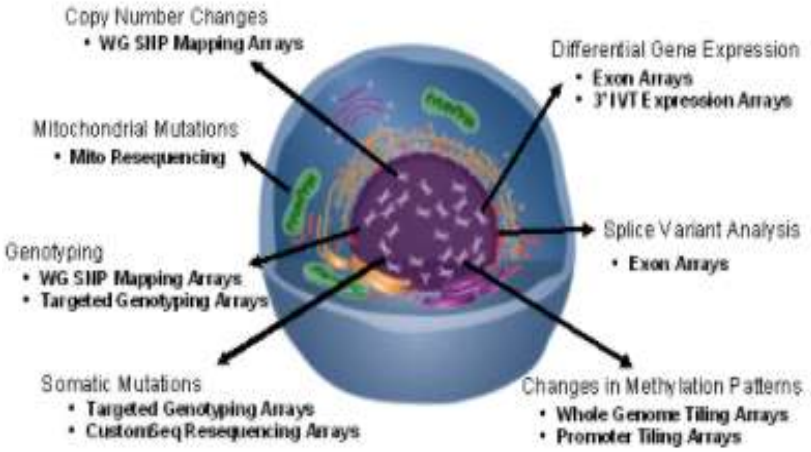
Affymetrix Instrumentation



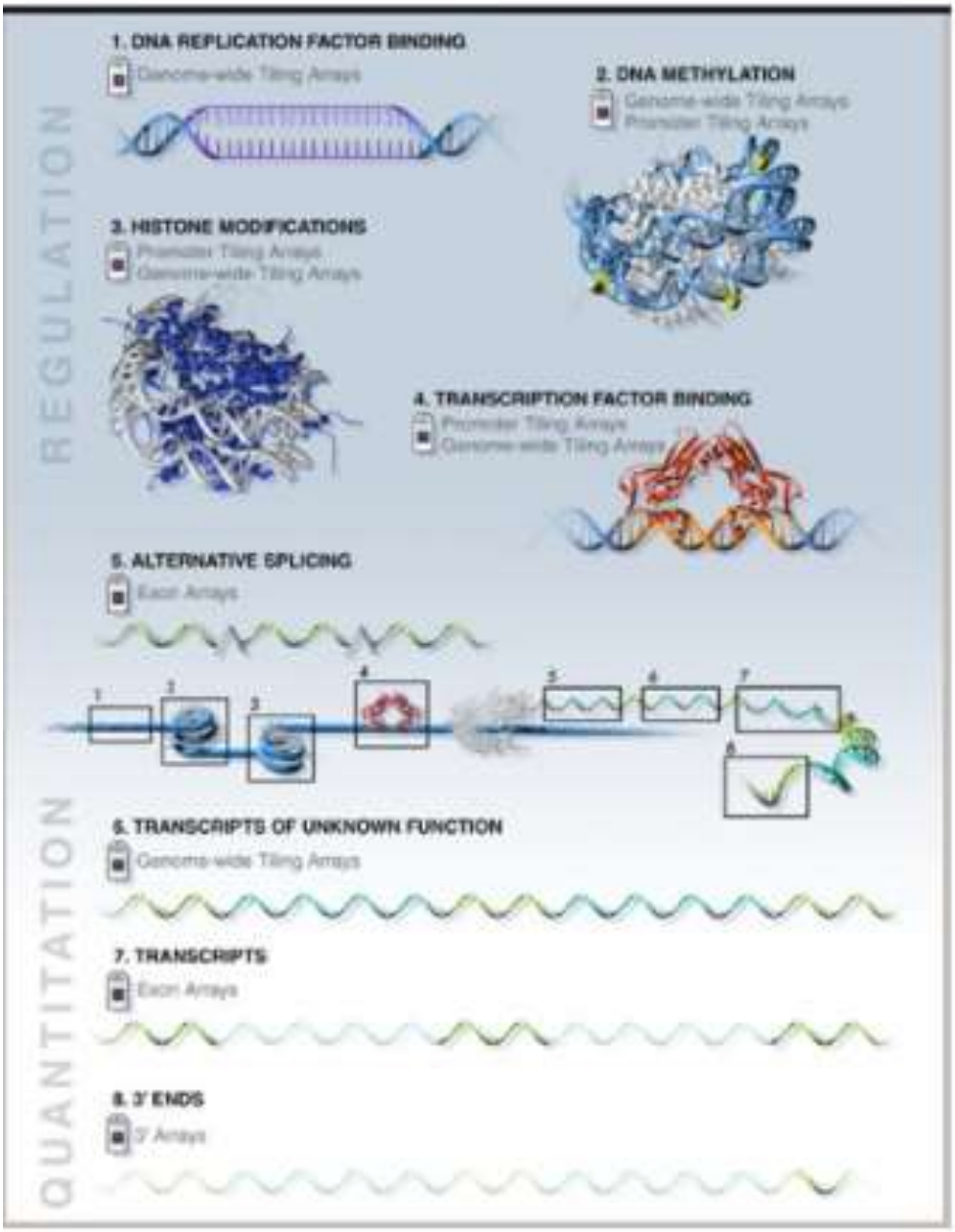
Chip features

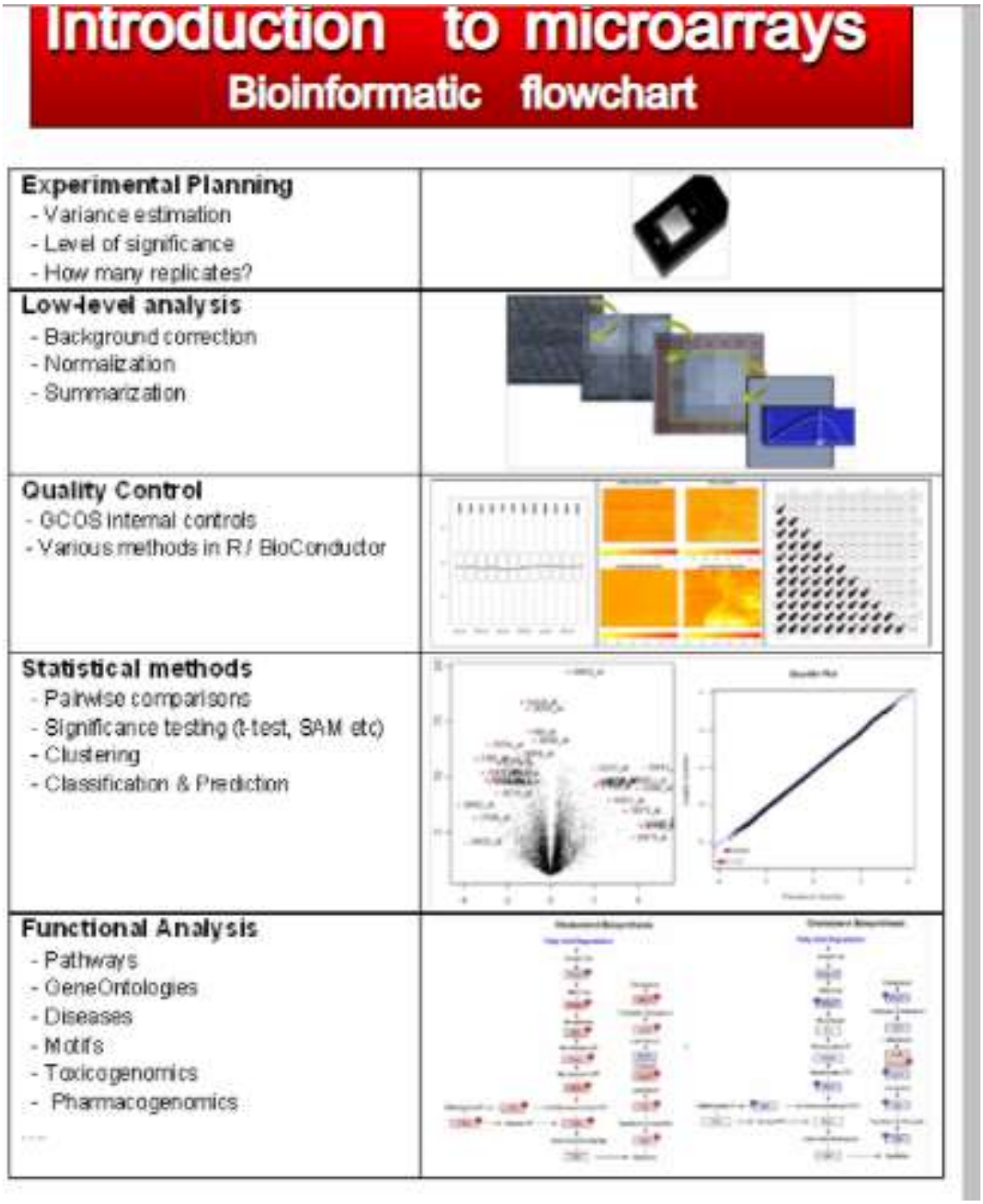


Arrays and Applications



Affymetrix Applications





Affymetrix GeneChip® System for Gene Expression Analysis

The Affymetrix GeneChip system is a commercial microarray platform that allows whole genome gene expression analysis for a wide variety of experimental organisms (http://www.affymetrix.com/products/arrays/index.affx).

This system has three major advantages over other array systems: it is easy to get rapid results; it has the capability to monitor the expression of every gene in the genome; and it is the most widely used commercial microarray platform.

Affymetrix GeneChip® miRNA Array

The GeneChip® miRNA Array is a powerful and cost effective tool for studying the role of microRNAs (miRNAs). The array provides comprehensive miRNA coverage (Sanger miRNA database V20 content and additional human small nuclear RNAs (snoRNAs and scaRNAs)) with multiple organisms (Human, mouse, rat, canine, and monkey) on a single array.

Input amounts: 0.13 – 3µg of total RNA or low molecular weight RNA enriched from 0.1-3 ug of total RNA (for new users we recommend to start with 1 ug of total RNA or LMW RNA enriched from 1 ug of total RNA).

Affymetrix microarray solutions are now branded Applied Biosystems and include all necessary components for a microarray experiment, from arrays and reagents to instruments and software. Our solutions enable scientists and clinicians to understand underlying disease mechanisms, identify biomarkers for personalized medicine, create novel molecular diagnostic tests, and improve genetic marker-assisted breeding programs in agriculture, thereby translating research results into biology for a better world.

Key applications

Transcriptome Analysis

Phenotypic abnormalities are rarely a result of expression changes in single genes, so generating a comprehensive expression profile is critical when studying normal biology and disease processes. Profile all known coding and non-coding splice variants.

Human Genotyping for Precision Medicine Research

Large-scale genotyping studies aimed at improving understanding disease risk and drug response are helping to pave the way toward precision medicine. To be successful, these studies require affordable, high-density genotyping arrays with accurate imputation and assurance that every marker will be on every array, every time.

Cytogenetics Analysis

Microarray-based assays provide a genome-wide approach that enables high-solution DNA copy number analysis to detect gains, losses, loss/absence of heterozygosity (LOH/AOH), copy-neutral LOH (cnLOH), regions identical-by-descent, and mosaicism in a single assay.

miRNA Profiling

Perform comprehensive miRNA profiling from as little as 130 ng and start exploring the role of miRNA in.24 hours—no bioinformatics resources required.

Large-scale Biobank Genotyping

Our Axiom Biobank Genotyping Arrays feature imputation-aware modular designs that enable scientists to conduct large-scale, state-of-the-art traits and population studies that help us understand how complex interactions between genes, environment, and lifestyle relate to health.

Plant and Animal Genotyping

Agrigenomics research is growing as climate change, population growth, and urbanization threaten farmers' ability to meet the world's food demands. To address these needs, breeders and farmers are employing new genomic strategies. Our powerful, flexible array-based genotyping solutions can help.

The goal of microarray image analysis is to extract intensity descriptors from each spot that represent gene expression levels and input features for further analysis. Biological conclusions are then drawn based on the results from data mining and statistical analysis of all extracted features.

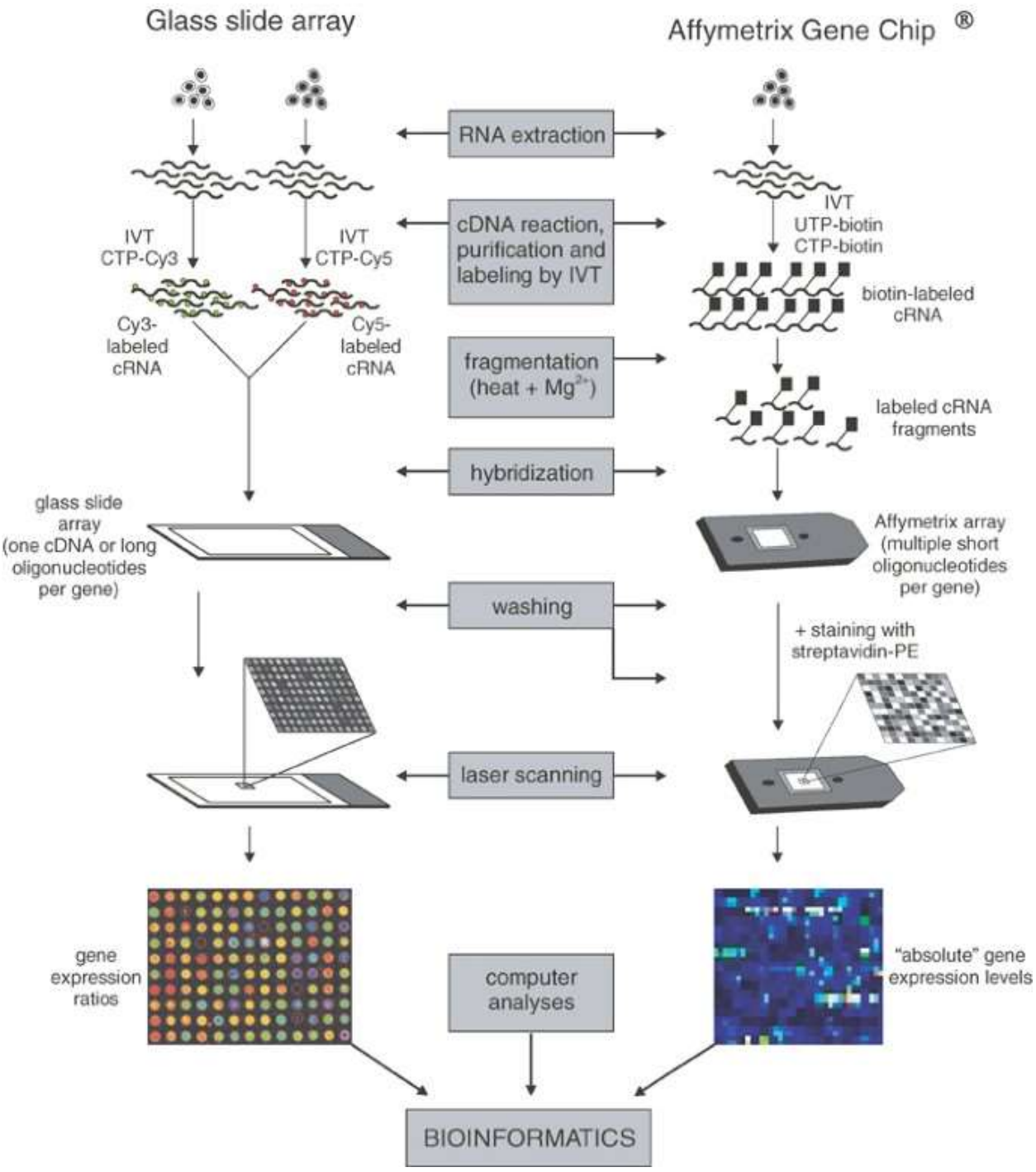
Components of DNA Microarray image analysis are (1) Grid Alignment Problem, (2) Foreground Separation, (3) Quality Assurance, (4) Quantification and (5) Normalization. Additionally, the data management must conform with the Minimal Information About Microarray Experiments (MIAME) standard.

Input: Laser image scans (data) and underlying experiment hypotheses or experiment designs (prior knowledge).

Output: Conclusions about statistical behavior of measurements and thus the the test of the hypotheses or knowledge. The results are derived automatically from data (machine learning perspective) for subsequent model fitting.

Applications of Microarrays

- Gene Discovery: DNA Microarray technology helps in the identification of new genes, know about their functioning and expression levels under different conditions.
- Disease Diagnosis: DNA Microarray technology helps researchers learn more about different diseases such as heart diseases, mental illness, infectious disease and especially the study of cancer. Until recently, different types of cancer have been classified on the basis of the organs in which the tumors develop. Now, with the evolution of microarray technology, it will be possible for the researchers to further classify the types of cancer on the basis of the patterns of gene activity in the tumor cells. This will tremendously help the pharmaceutical community to develop more effective drugs as the treatment strategies will be targeted directly to the specific type of cancer.
- Drug Discovery: Microarray technology has extensive application in Pharmacogenomics. Pharmacogenomics is the study of correlations between therapeutic responses to drugs and the genetic profiles of the patients. Comparative analysis of the genes from a diseased and a normal cell will help the identification of the biochemical constitution of the proteins synthesized by the diseased genes. The researchers can use this information to synthesize drugs which combat with these proteins and reduce their effect.
- Toxicological Research: Microarray technology provides a robust platform for the research of the impact of toxins on the cells and their passing on to the progeny. Toxicogenomics establishes correlation between responses to toxicants and the changes in the genetic profiles of the cells exposed to such toxicants.


Glass slide array
Affymetrix Gene Chip ®
RNA extraction
cDNA reaction, purification and labeling by IVT
fragmentation (heat + Mg2+)
hybridization
washing
laser scanning
IVT CTP-Cy3
IVT CTP-Cy5
Cy3-labeled cRNA
Cy5-labeled cRNA
IVT UTP-biotin CTP-biotin
biotin-labeled cRNA
labeled cRNA fragments
glass slide array (one cDNA or long oligonucleotides per gene)
Affymetrix array (multiple short oligonucleotides per gene)
+ staining with streptavidin-PE
gene expression ratios
computer analyses
"absolute" gene expression levels
BIOINFORMATICS

UNIT – 2- SBIA5304-MICROARRAY DATAANALYSIS

UNIT II IMAGE PROCESSING

Image processing, feature extraction, identifying positions of features- Normalization – data cleaning and transformation, within array normalization, between array normalization, measuring and quantifying microarray variability –variability between replicate features on an array-, variability between hybridizations to different arrays. Analysis of differentially expressed genes- significance analysis of microarrays.

Microarray Image Analysis

- Microarray image processing leads to the characterization of gene expression levels simultaneously, for all cellular transcripts (mRNAs) in a single experiment.
- The calculation of expression levels for each microarray spot/gene is a crucial step to extract valuable information.
- By measuring the mRNA levels for the whole genome, the microarray experiments are capable to study functionality, pathological phenotype, and response of cells to a pharmaceutical treatment. The processing of the extensive number of non-homogeneous data contained in microarray images is still a challenge.

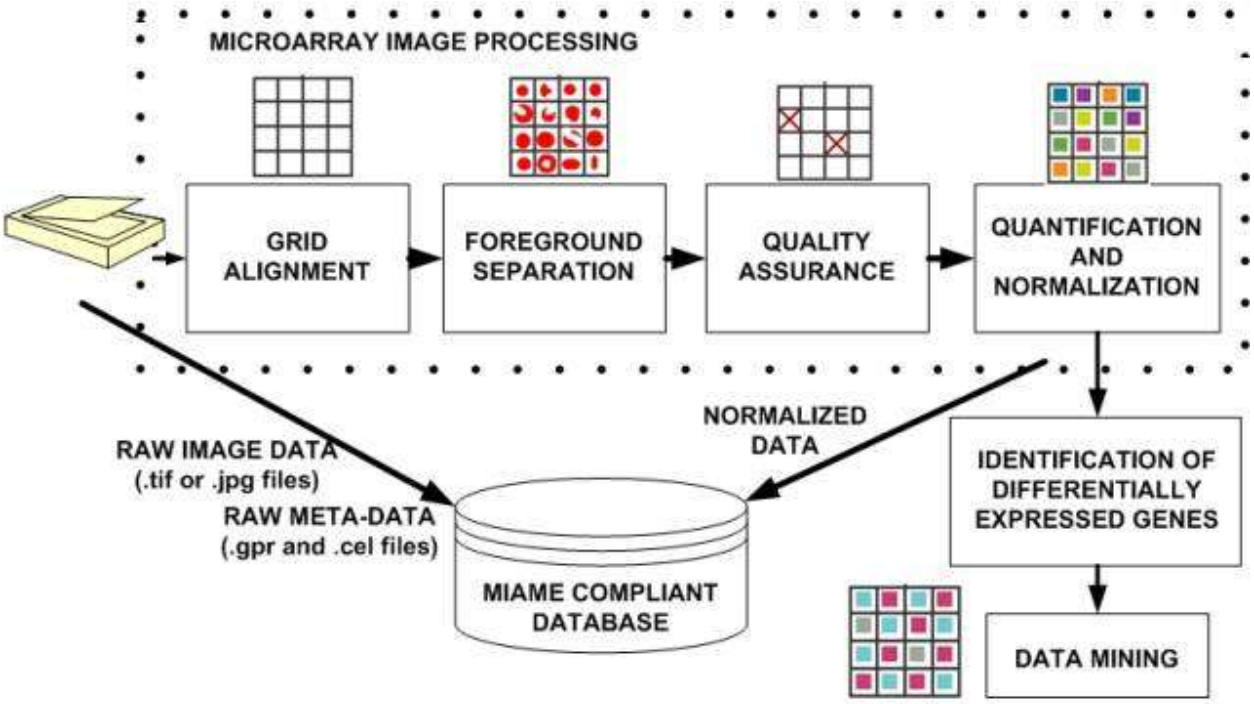
INTRODUCTION

Image analysis is an important aspect of microarray experiments. It can have a potentially large impact on subsequent analysis such as clustering or the identification of differentially expressed genes. In microarray experiments, hybridised arrays are imaged in a microarray scanner to produce red and green fluorescence intensity measurements at each of a large collection of pixels which together cover the array. These fluorescence intensities correspond to the levels of hybridisation of the two samples to the DNA sequences spotted on the slide. Fluorescence intensities are usually stored as 16-bit images which we view as 'raw' data.

microarray images can generally be separated into three tasks.

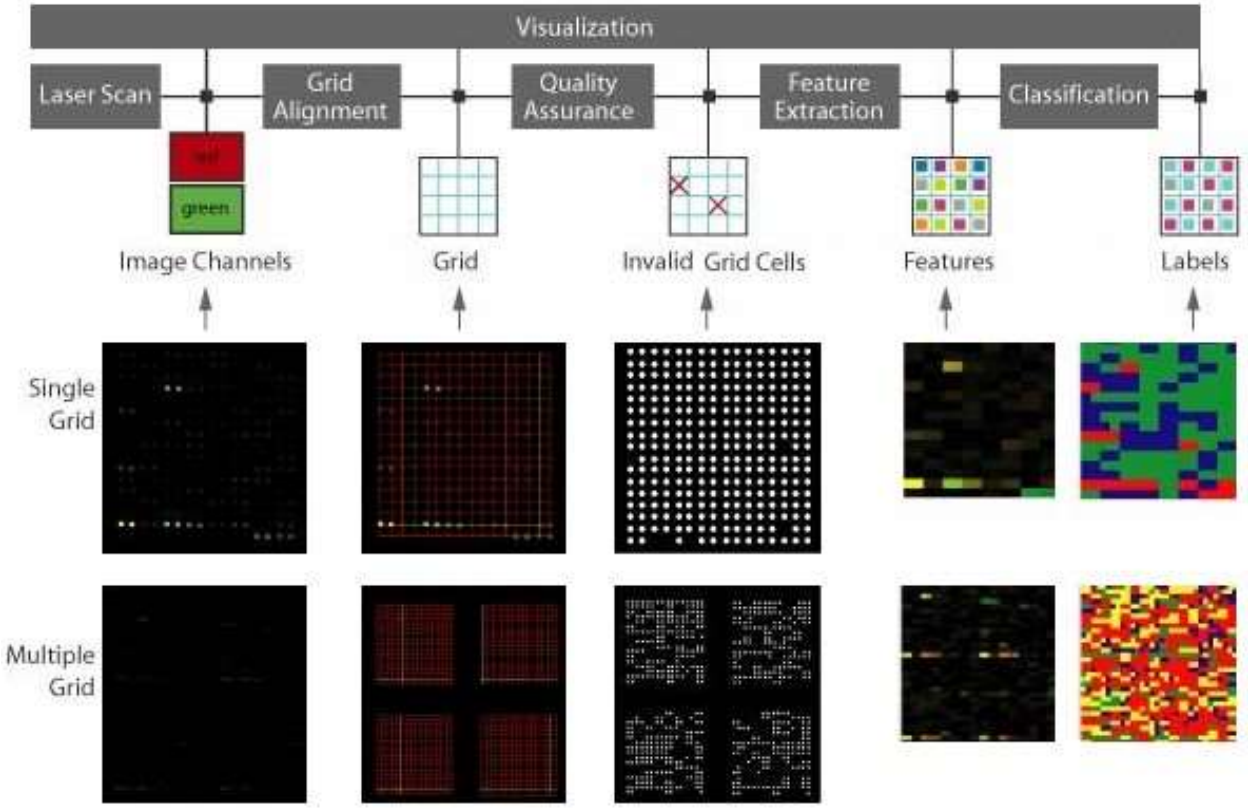
- *Addressing* or *gridding* is the process of assigning coordinates to each of the spots. Automating this part of the procedure permits high-throughput analysis.
- *Segmentation* allows the classification of pixels either as foreground – that is, within printed DNA spot – or as background.
- The *intensity extraction* step includes calculating, for each spot on the array, red and green foreground fluorescence intensity pairs (R,G), background intensities and, possibly, quality measures.

- Typically, the microarray images are stored in the Tagged Image File Format (TIFF) as a two-dimensional array of intensities.
- In a two colour microarray experiment, two microarray images are available, each image being recorded from a specific cyanine dye.
- The images are denoted by ICy3 and ICy5, corresponding to Cy3 and Cy5 dyes, respectively.



- Microarray grid alignment and foreground separation are the basic processing steps of DNA microarray images that affect the quality of gene expression information, and hence impact our confidence in any data-derived biological conclusions.
- Thus, understanding microarray data processing steps becomes critical for performing optimal microarray data analysis.
- The workflow of microarray data processing starts with raw image data acquired with laser scanners and ends with the results of data mining that have to be interpreted by biologists.
- The microarray data processing workflow includes issues related to
- data management (e.g., MIAME compliant database,
- (2) image processing (grid alignment, foreground separation, spot quality assessment, data quantification and normalization,

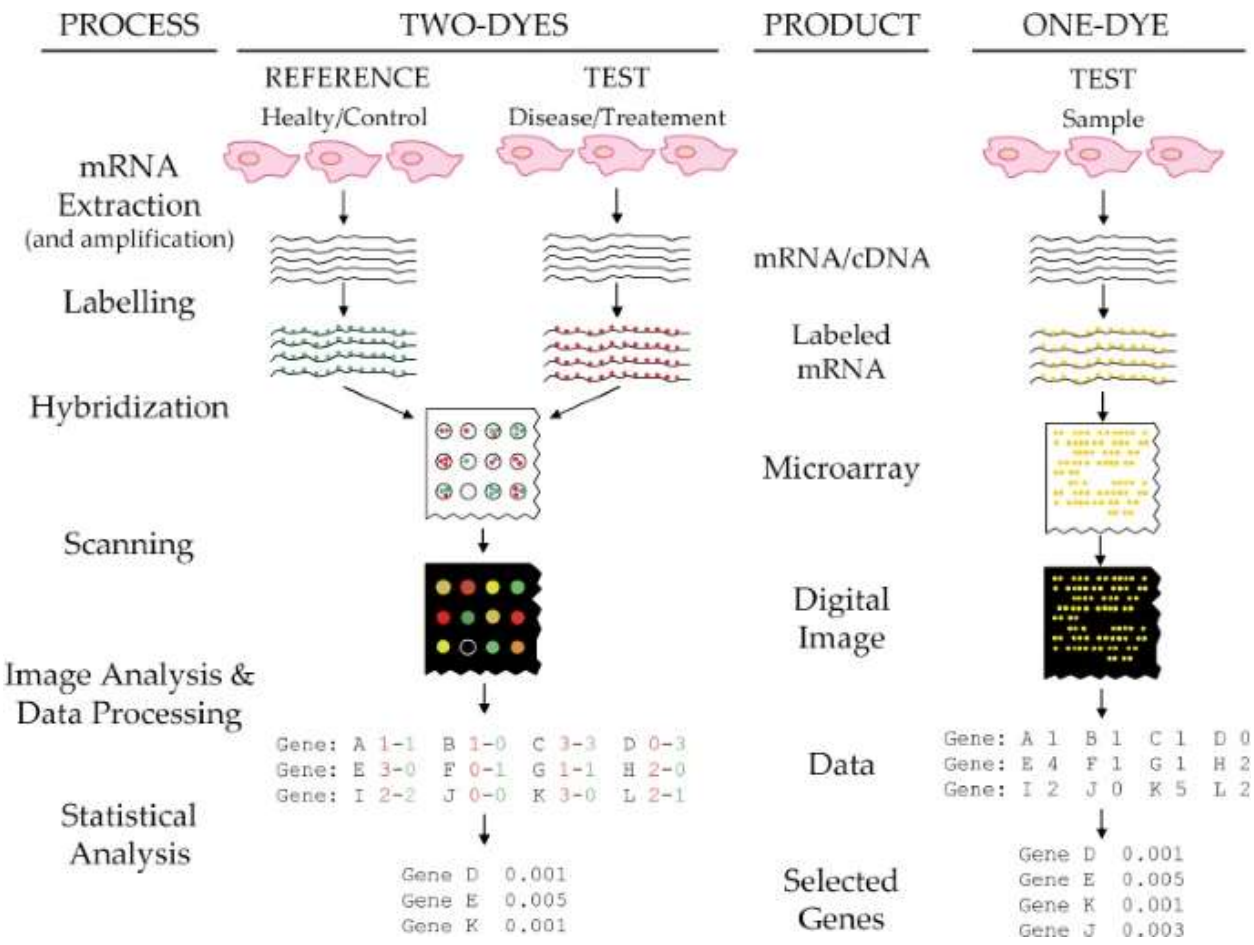
- (3) data analysis (identification of differentially expressed genes, data mining, integration with other knowledge sources, and quality and repeatability assessments of results, and
- (4) biological interpretation (visualization).
- **The main objective of this project is related to image processing, namely grid alignment, foreground separation, spot quality assessment, data quantification, normalization and visualization.**

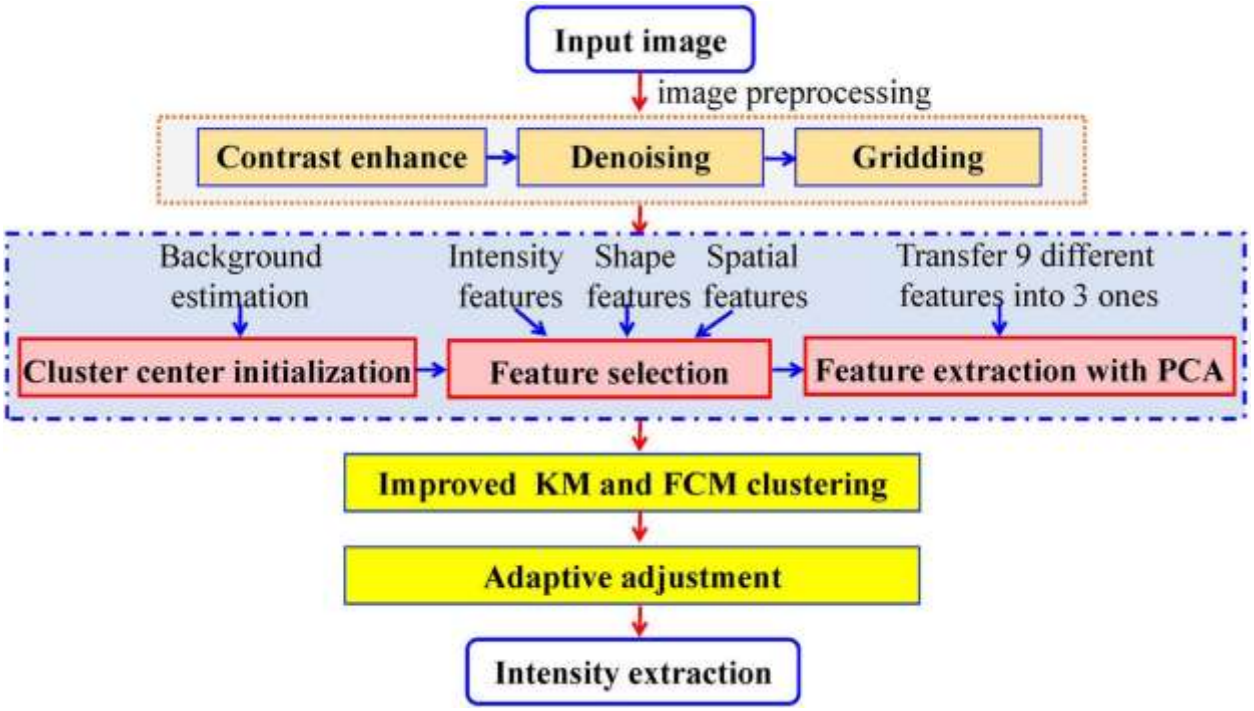


- Microarray data processing workflow: Fluorescent DNA microarray images obtained from laser scanners containing a 2D array of dots with two channels of 532nm (red) and 632nm (green) wavelengths.
- The grid alignment is performed producing a set of lines intersecting at each dot.
- Dots define a valid foreground.
- Quality assurance screening eliminates grid cells with unreliable microarray information.
- Finally, image of sample mean values extracted at each grid cell using particular mask is extracted and colored in a red-green-blue space with color assigned to each cluster/pixel.

- Statistics of each cluster can be viewed in the text area
- Microarray images represent a collection of microarray spots arranged in one or more sub-grids, each grid representing a two dimensional array of spots.
- Image processing technique are used further on in order to determine spot location within each subgrid, spot sizes, spot intensities and background intensities values which are typically delivered as raw data parameters for microarray image analysis and interpretation
- **A typical microarray image is generated from an array of cDNA probes which is hybridized to two samples, one being red fluor-tagged and the other green fluor-tagged. The composite color image is constructed by placing each monochrome image into the appropriate color channel.**
- **The tasks of microarray image analysis can be further-divided into following tasks:**

1. ***Array target segmentation***
2. ***Background intensity extraction***
3. ***Target detection***
4. ***Target intensity extraction***
5. ***Normalization and Ratio analysis***
6. ***Measurement quality assessment***
7. ***software package based interpretation***





Segmentation is the method of segregating a spitting an image into multiple fundamental fragments. The

segmentation phase of the image study shows a key role in the statistical analysis, a step where the data is produced.

Four categories of methods for microarray image segmentation are

(a) fixed /adaptive circle segmentation

(b) Histogram based techniques

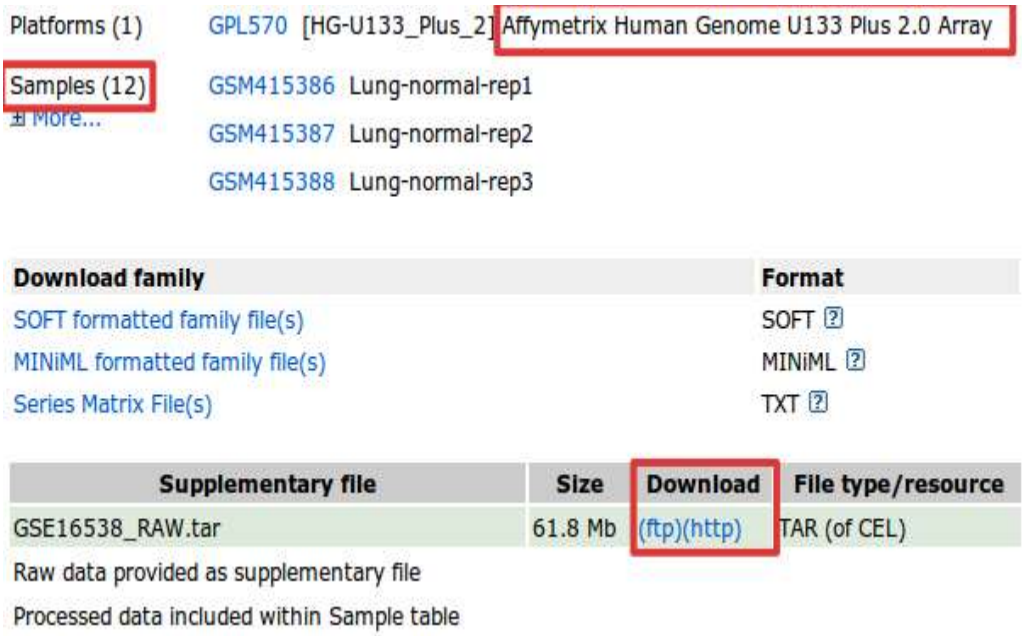
(c) adaptive shape segmentation

(d) Machine learning techniques.

Thus the integration of machine learning in Image processing will contribute a better analysis of medical and biological data

- The fourth category is based on machine learning techniques.
- There are two categories in this method.
- They are
- (a) supervised segmentation techniques
- (b) unsupervised segmentation technique.
- More specifically, methods in unsupervised category employ clustering algorithms, such as k-means, hybrid k-means, fuzzy c-means,expectation-maximization and partitioning method for segmentation of microarray images.
- An important first step of any microarray experiment is the normalization of the samples.
- Although the relative impacts differ from platform to platform and sample preparation, non-biological differences in microarray signals can stem from a variety of factors, such as: global constant background noise, non-specific binding signal, non-linear signal response between samples, bad spots on the chip due to dust or bubbles or rare manufacturing defects, labeling efficiency, hybridization efficiency, and RNA quality.
- Statistical analysis of microarray data is started through software programs using CEL files defined as raw data.
- Prior to the start of the analysis, quality assessment of raw data is performed as the first step.

- In order to evaluate the homogeneity of the arrays and to compare the density distribution between the arrays, box graphs are plotted for each array using the densities of the logarithm2 base of the raw data.
- Images of the CEL files are obtained to observe the dimensional distributions of the densities on each array and to detect dimensional artifacts.
- CEL file: Cell intensity file, probe level PM and MM values.

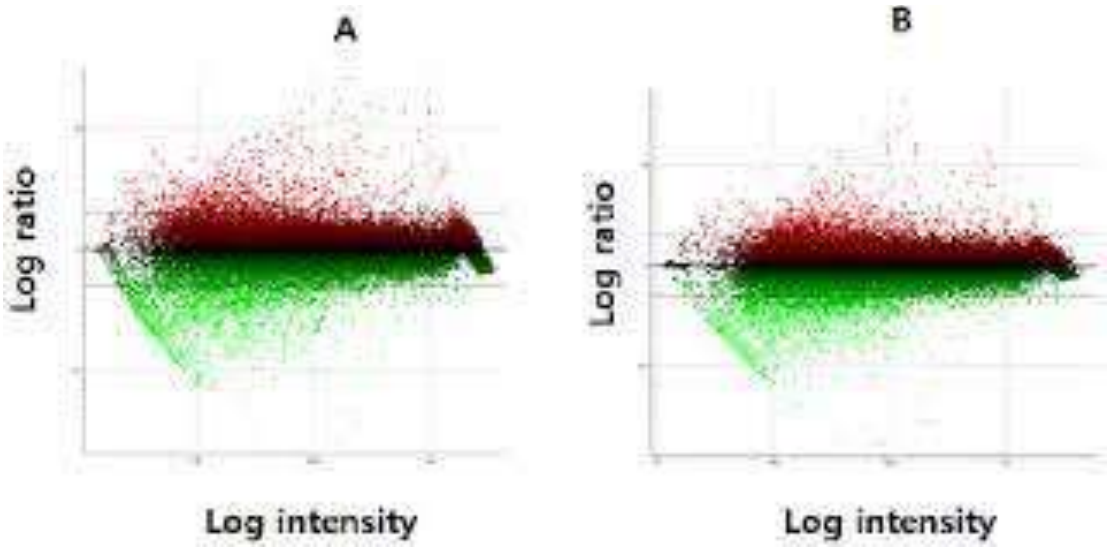


Each gene or portion of a gene is represented by 1 to 20 oligonucleotides of 25 base-pairs.

Probe: an oligonucleotide of 25 base-pairs, i.e., a 25-mer

- Perfect match (PM): A 25-mer complementary to a reference sequence of interest (e.g., part of a gene).
- Mismatch (MM): same as PM but with a single base change for the middle (13th) base (transversion purine <-> pyrimidine, G <->C, A <->T). Used to measure non-specific binding and background noise.
- Probe-pair: a (PM,MM) pair. Probe-pair set: a collection of probe-pairs (1 to 20) related to a common gene or fraction of a gene.
- Affy ID: an identifier for a probe-pair set.

- MA-plots are used to compare the expression values for all possible pair of arrays with a probeset-wise median array.
- The MA plots are generated by plotting M values which are obtained by logarithmic ratios versus A values which are average logarithmic intensity values.
- The pre-normalization quality control step can be complemented by histograms drawn to assess the density distributions of each array



- After quality control of raw data, background correction and normalization should be applied to the data using background correction methods such as RMA (Robust Multiple-Array Average) method.
- With the RMA method, the probe-level signal is removed from the background signal.
- Quantile normalization is performed by the RMA method and it is ensured that all the arrays have the same quantile.
- Using the RMA method, the expression set to be used in the analysis is generated by normalized and the background corrected intensities.
- After the background correction and the normalization methods are performed, box charts related to each array are drawn to re-evaluate the quality control.

- Following normalization and background correction, a list of genes that differ between two different conditions can be obtained by applying various statistical tests to the expression dataset to be used for analysis

Preprocessing of microarray data

- Measurement values may have undergone various adjustments in the device system, such as calibration.
- Thus, in the presentation of gene expression data, it must be explained how the values are generated by the device system.
- These expression measures always contain a component called **"background noise."**
- Local background noise levels are measured from the areas of the glass slide that do not contain probes.
- The background correction tries to remove non-specific background noise and local variations of the overall signal level on each chip.
- The most common method to remove the background effect is to remove the measured fluorescence intensity around the spots.

Microarray gene expression data sets consist of ω_{gn} gene expression values, with $g = 1, \ldots, G$ genes and $n = 1, \ldots, N$ samples. ω_{gn} values are arranged in a $G \times N$ data matrix, where each gene corresponds to one row and each sample to one column. The readout gene expression value ω_{gn} can be statistically defined as the sum of the true gene expression value x_{gn} and the background noise B_{gn} components [6];

$$\omega_{gn} = x_{gn} + B_{gn} \qquad (1)$$

The structure and correction of the background noise depend on the microarray technology used. Spot array data provides an estimate of background noise B_{gn}, with uncorrected expression intensities ω_{gn} values. If the background estimate is expressed as $\hat{B}_{gn}$, background corrected expression value, $\omega_{gn}^{(c)}$ is given as follows[6];

$$\omega_{gn}^{(c)} = \omega_{gn} - \hat{B}_{gn} = \left(x_{gn} + B_{gn}\right) - \hat{B}_{gn} \qquad (2)$$

The most common methods used for background correction in microarray analysis are; The “Robust Multi-Array Average (RMA) Background Correction” method and the “MAS 5.0 Background Extraction” methods

RMA background correction: RMA background correction is a method that uses only Perfect Match (PM) intensities. PM values are corrected using a global model for the distribution of probe intensities [7].

The model is based on the experimental distribution of probe intensities. Observed PM probes are modeled as a Gaussian noise component with μ average and σ^2 variance [7].

To avoid negative expression values, the normal distribution is truncated at zero. If the observed density is assumed to be *Y*, the correction will be as follows;

$$E\left(S\middle|Y=y\right)=a+b\frac{\varnothing\left(\frac{a}{b}\right)-\varnothing\left(\frac{y-a}{b}\right)}{\Phi\left(\frac{a}{b}\right)+\Phi\left(\frac{y-a}{b}\right)-1} \quad (3)$$

$\alpha = s-\mu-\sigma^2\alpha$ and $b=\sigma$ where *S* is an averaged exponential signal component with α mean. ϕ and Φ are the standard normal density and distribution functions, respectively [7].

MAS 5.0 background correction: In the MAS 5.0 background correction method, the chip is divided into a rectangular grid with k rectangular regions. At each region, at least 2% of the probe intensities are used to calculate a background value for this grid. Then, each probe intensity is corrected based on a weighted average of the background values. The weights depend on the distance between the probes and the center of gravity of the grid [7].

Weights are calculated as follows;

$$\omega_k(x,y) = \frac{1}{d_k^2(x,y)+s_0} \quad (4)$$

Where $d_k(x,y)$ is a Euclidean distance from (x,y) position to the center of gravity of region k and s_0 is correction coefficient.

In MAS 5.0 Background Correction method, both Perfect Match (PM) and Mismatch (MM) probes are corrected [7].

- RMA background correction has been one of the most commonly used pre-processing method in the recent literature.
- Performed assessments of the measure's precision, consistency of fold change, and specificity and sensitivity of the measure's ability to detect differential expression and demonstrated the substantial benefits of using the RMA measure to users of the Gene Chip technology.
- They used data from spike-in and dilution experiments to conduct various assessments on the MAS 5.0, dChip and RMA expression measures.
- Irizarry have demonstrated that RMA has similar accuracy but better precision than the other two summaries and RMA provides more consistent estimates of fold change

Quality assessment:

- It is necessary to evaluate the quality of the data before the normalization of the arrays. Quality control assessment should be carried out to determine whether the quality of experimental data is acceptable and whether any hybridization should be repeated.

- Various descriptive data plots are drawn to identify potential problems with hybridization or other experimental structures in the quality control evaluation process.
- Quality control plots are basically divided into diagnostic and spot statistics .
- **Diagnostic Plots:** The diagnostic plots include various plots such as MA-plots for evaluating intensity bias and histograms for examining signal-to-noise ratios for each channel.
- **Diagnostic plots** are usually used to observe non-linear trends between two channels

a. MA plots: *M* and *A* are commonly used variables in the analysis of two-color arrays. *A* is defined as follows;

$$A = log_2\sqrt{Cy5 \cdot Cy3} = \frac{1}{2}\left[log_2(Cy5) + log_2(Cy3)\right] \qquad (5)$$

Cy5 and Cy3 denote green and red dye intensities for a given spot, respectively. *A* variable is a measure of the total intensity of the logarithmic transformation of a spot. Thus, if the combined red and green intensities are high for a particular spot, the *A* value will also be high [7,9].

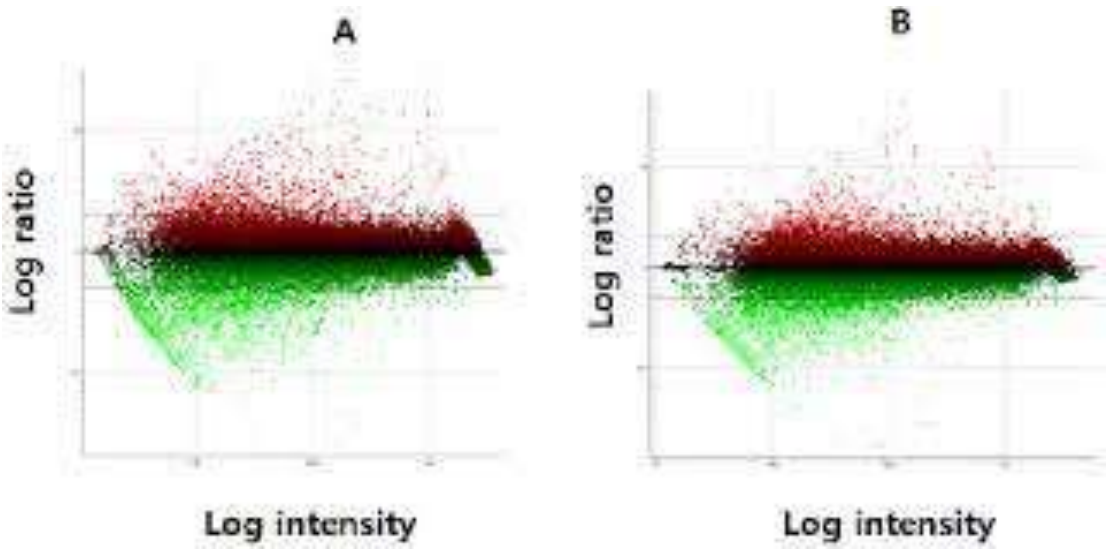
M variable is defined as follows;

$$M = log_2\frac{Cy5}{Cy3} = log_2(Cy5) - log_2(Cy3) \qquad (6)$$

The *M* variable is the logarithmic transformation of the intensity ratio. The *M* value shows which of the red and green dyes are more binding to a particular spot array [7].

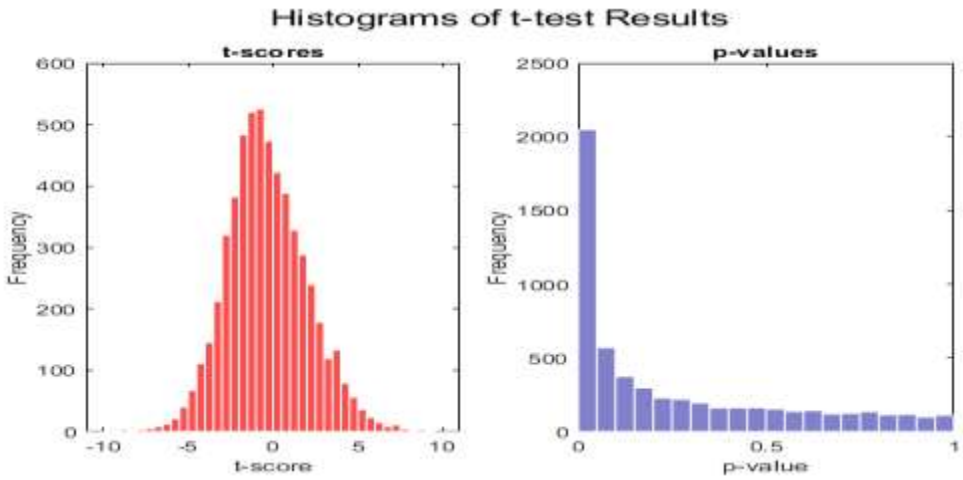
MA plots are used to investigate density bias. *A* disproportionate amount of spot above or below the x-axis on the graph indicates a problem in the array. MA plots are an indication of whether normalization within the array is required [6,7].

Alvord et al.demonstrated the use of some of the exploratory plots including boxplots, volcano plots and MA plots for the expression level data on the soybean genome [14]. Lu et al.'s study, can be cited as an example of MA-plot application in method comparison studies, in which MA-plots were created on the raw data and normalized data to compare normalization methods [15].



Histograms:

- In microarray designs, it is very important to obtain the histograms of the p-values of tests conducted to identify different gene expression.
- Histograms are graphs that are easy to interpret and contain considerable information.
- A histogram is an indication of whether there is a signal in the gene and whether the genes are differently expressed. Histograms also allow for estimation of how many genes are differentially expressed in reality.



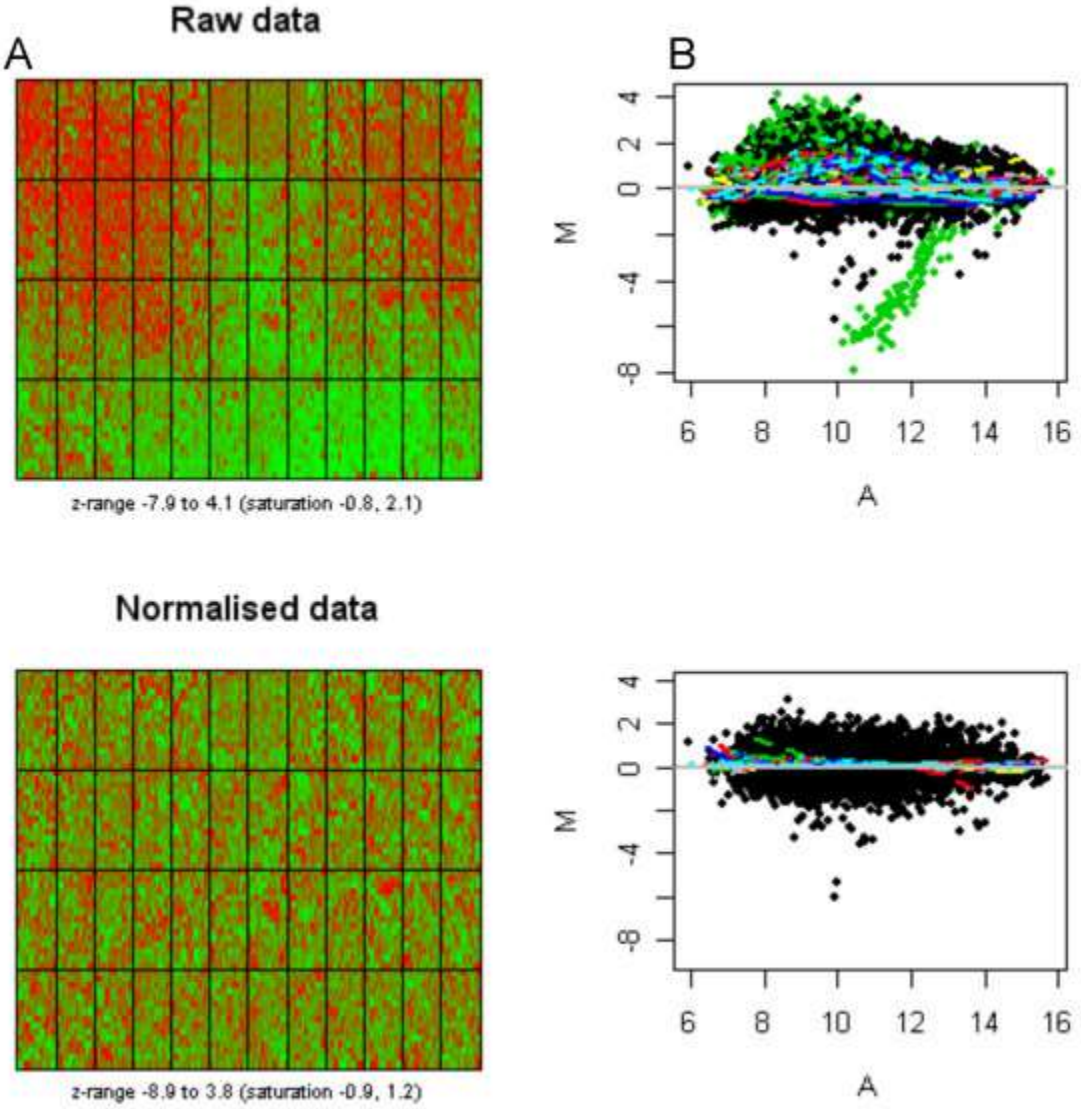
Spot statistics plots:

- Spot Statistics help to predict the structures of spot and hybridizations.

- The main plots that can be obtained with spot statistics are spatial plots, box plots, scatter diagrams and volcanic plots .

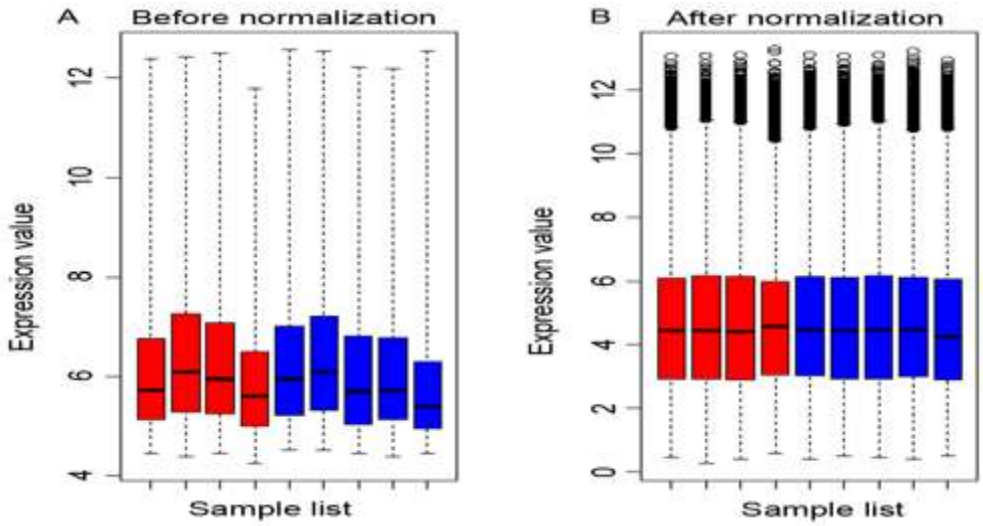
Spatial plots:

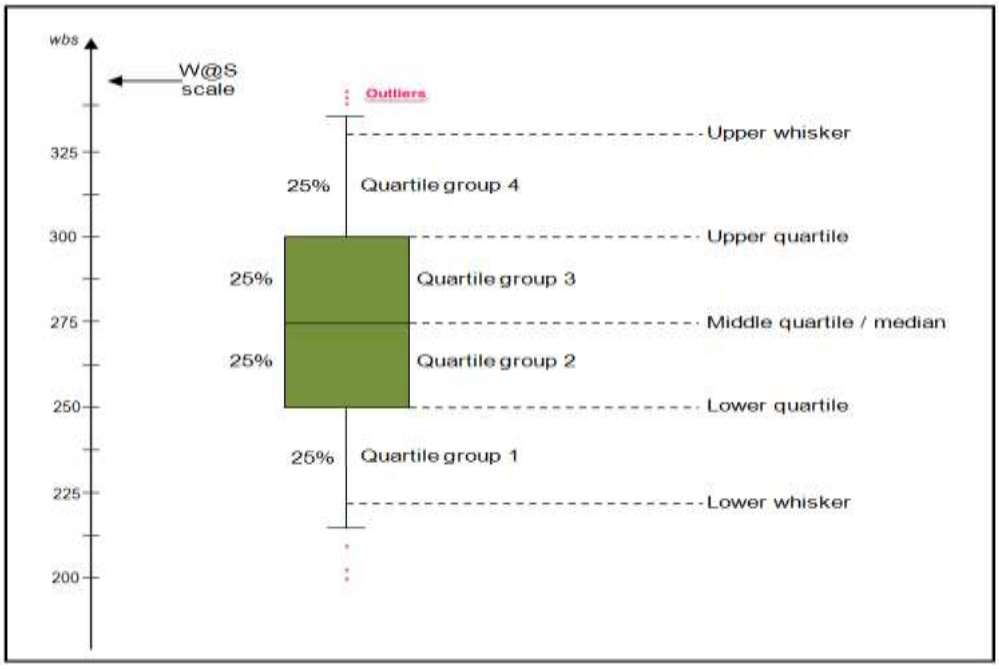
- Spatial plots are used to reveal irregular spot and hybridization structures.
- Spatial plots are used to observe the spatial distributions of the intensities on each array and to detect the artifacts.
- Spatial plots play a fundamental role in determining the background correction, depending on whether there are any dimensional artifacts on the arrays



Box plots:

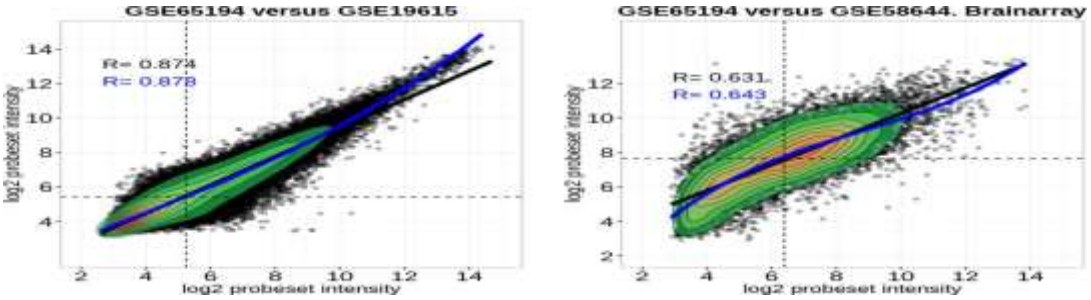
Box plots are one of the most commonly used plots for displaying spot and hybridization structures. At the same time, box plots can be drawn to understand the scale differences between different arrays. It is necessary to evaluate the box plots to see if between-array normalization is required. The homogeneity of the arrays can be observed quite clearly from the box plots .





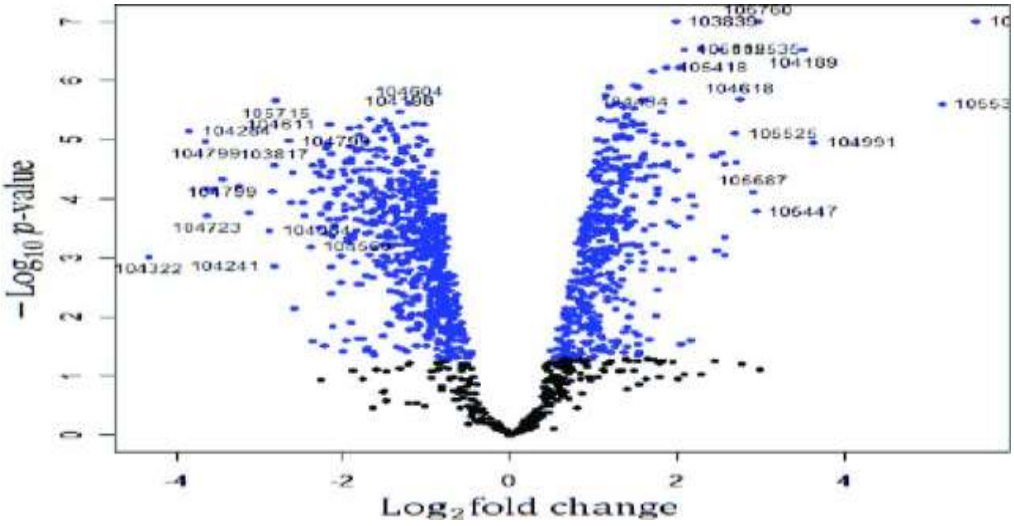
Scatter diagrams:

- Scatter diagrams used to compare the expression values of two samples are the most commonly used plots for visualizing microarray data.
- In the first step of the microarray data analysis, a scatter diagram is drawn between the two intensity channels to view the general structures and variability.
- Scatter diagrams are also commonly used to find out slides lying away from the center, which have abnormal expression structures .



Volcano plots:

- Volcano Plots are used to summarize fold change and t-test criteria.
- A volcano plot is constructed by plotting the negative log of the p-value on the y-axis and log of the fold change between the two conditions on the x-axis.
- For each gene, there is a point on the graph that represents the t-statistic and the fold change



- *A volcano plot of the genes in microarray. The Log 2 fold changes and their corresponding-log 10 p-value of all genes were taken for construction of the volcano plot in the microarray. The genes*

with p < 0.05 are depicted in blue dots. All other genes that were not found to be significant altered are in black dots in this array

- **Normalization**
- The aim of normalization methods for large scale expression data, including microarray and RNA-seq, is to eliminate systematic experimental bias and technical variation while preserving biological variation.
- Dozens of normalization methods for correcting non-linear experimental differences between arrays have been developed during the last two decades. Among them, quantile and lowess are well-adopted for analyzing microarray expression data.
- **Expression ratios**: the primary comparison
- Most microarray experiments investigate relationships between
- related biological samples based on patterns of expression, and
- the simplest approach looks for genes that are differentially
- expressed.
- If we have an array that **has N array distinct elements,**and compare a query and a reference sample, which for convenience we will call R and G, respectively (for the red and green colors commonly used to represent array data), then the ratio (T) for the ith gene (where i is an index running over all the arrayed genes from 1 to **Narr**ay) can be written as

$$T_i = \frac{R_i}{G_i},$$

Normalization:

- The purpose of the normalization phase is to adjust the data according to the technical variation. Variations can cause measurement differences between general fluorescence intensity levels of various arrays. The normalization process is necessary to make the measured values obtained from different arrays comparable.

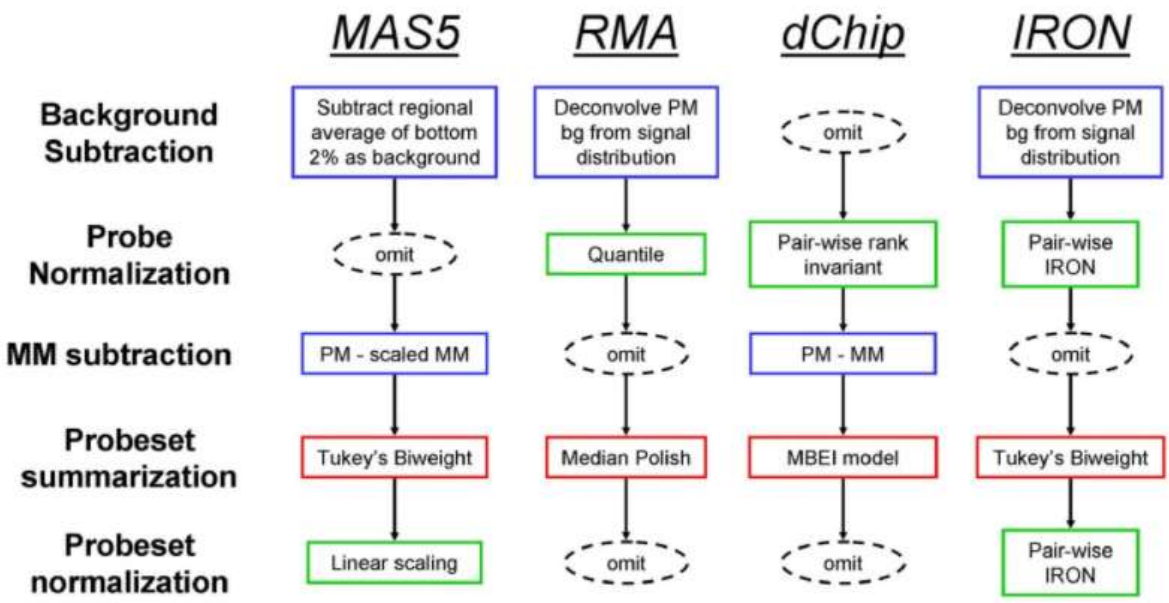
- Normalization methods depend on which microarray technology is used. Generally, logarithmically transformed data are used for further analysis.
- The most commonly used methods of normalization are as follows

1. Scaling Normalization Method
2. Nonlinear Normalization Methods
3. Quantile Normalization
4. Cyclic Loess Normalization
5. Contrast Normalization

- Normalization
- Typically, the first transformation applied to expression data,referred to as normalization, adjusts the individual hybridization.
- (Note that this definition does not limit us to any particular array technology: the measures Ri and Gi can be made on either a single array or on two replicate arrays. Furthermore, all the transformations described below can be applied to data from any microarray platform.)
- Although ratios provide an intuitive measure of expression changes, they have the disadvantage of treating up- and downregulated genes differently. Genes upregulated by a factor of 2 have an expression ratio of 2, whereas those downregulated by the same factor have an expression ratio of (–0.5).
- The most widely used alternative transformation of the ratio is the logarithm base 2, which has the advantage of producing a continuous spectrum of values and treating up- and downregulated genes in a similar fashion.
- Recall that logarithms treat numbers and their reciprocals symmetrically: ***log2(1) = 0, log2(2) = 1, log2(1/ 2) = −1, log2(4) = 2, log2(1/ 4) = −2, and so on.***
- The logarithms of the expression ratios are also treated symmetrically, so that a gene upregulated by a factor of 2 has a log2(ratio) of 1, a gene downregulated by a factor of 2 has a log2(ratio) of −1, and a gene expressed at a constant level (with a ratio of 1) has a log2(ratio) equal to zero.
- For the remainder of this discussion, log2(ratio) will be used to represent expression levels
- **There are three major normalization methods that are commonly employed**:

- linear scaling (MAS5), quantile normalization (RMA), and pair-wise rank-invariant normalization (dChip).
- Linear normalization is the simplest of the methods, which applies a global scaling factor to each chip (at the probeset level in MAS5) in order to scale all chips to the same trimmed mean intensity.
- Quantile normalization ranks the intensities for each chip, then replaces the intensities at each rank with the mean intensity for all probes of that rank across all chips, effecting a non-linear rank-dependent normalization.
- Pair-wise rank-invariant normalization normalizes all chips against a single reference chip by identifying a different subset of rank-invariant genes for each sample/reference chip pair, fitting a curve through the training set, then adjusting the intensities of the target chip in an intensity-dependent manner so that the fit curve will lie on the sample vs. reference diagonal of the scatterplot.
- ***Linear normalization*** is unable to correct for non-linear, intensity-dependent differences in gene expression between chips, but can be applied to a single chip, independently of other chips.
- ***Quantile normalization assumes*** that differential gene expression is symmetric, in that there will be a roughly equal number of up and down regulated genes with equal magnitude distributions. Due to its population-based signal, it requires a moderately large number of chips in order to work well, and may introduce unexpected artifacts, particularly in outlier samples, in small experiments, or experiments in which different cell/tissue types are represented.
- ***Rank-invariant normalization*** makes similar assumptions to those of quantile normalization, since both are rank based, but can be applied to as few as two chips.
- The three most commonly used software packages for processing Affymetrix microarrays, as evidenced by recently querying the GEO and ArrayExpress microarray repositories, are: RMA, MAS5 , and dChip
- Each of these employs different methods for background subtraction, signal normalization, and **probeset summarization** (an issue unique to Affymetrix arrays, where multiple probes for the same transcript are condensed into a single representative signal).
- **MAS5.0**

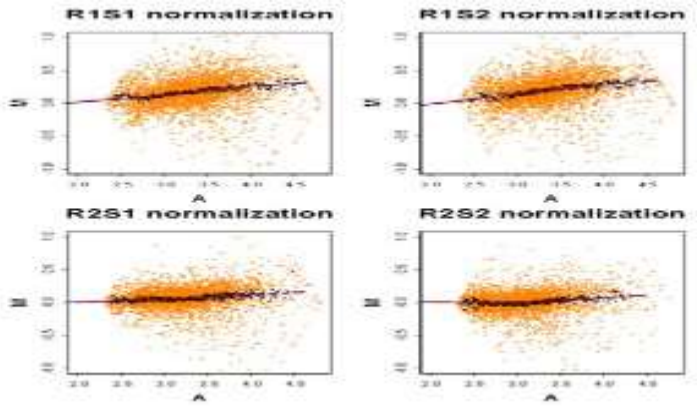
- A significant challenge with Affymetrix expression data is to provide an algorithm that combines the signals from the **multiple Perfect-Match (PM) and Mismatch (MM) probes** that target each transcript into a single value that sensitively and accurately represents its concentration.
- MAS5.0 does this by calculating a robust average of the (logged) PM-MM values; increased variation is observed at low signal strengths and is at least in part due to the extra noise generated by subtracting the MM values from their PM partners.
- Many gene expression normalization algorithms exist for Affymetrix GeneChip microarrays.
- The most popular of these is RMA, primarily due to the precision and low noise produced during the process.
- A significant strength of this and similar approaches is the use of the entire set of arrays during both normalization and model-based estimation of signal.



Comparison of MAS5, RMA, dChip, and IRON microarray post-processing pipelines. IRON combines components of both MAS5 and RMA, substituting a novel pair-wise iterative rank-order normalization method for normalization steps.

- Quantile normalization is an important normalization technique commonly used in high-dimensional data analysis.
- However, it is susceptible to class-effect proportion effects (the proportion of class-correlated variables in a dataset) and batch effects (the presence of potentially confounding technical variation) when applied blindly on whole data sets, resulting in higher false-positive and false-negative rates.

- **Pair-wise rank-invariant normalization (dchip)**
- Pair-wise rank-invariant normalization normalizes all chips against a single reference chip by identifying a different subset of rank-invariant genes for each sample/reference chip pair, fitting a curve through the training set, then adjusting the intensities of the target chip in an intensity-dependent manner so that the fit curve will lie on the sample vs. reference diagonal of the scatterplot.



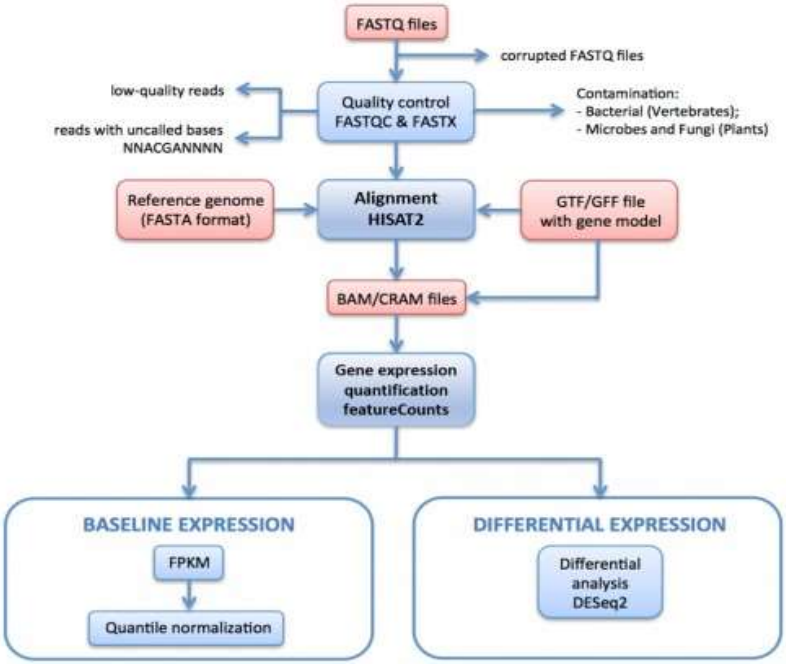
-

Three Major Approaches to Between Array Normalization

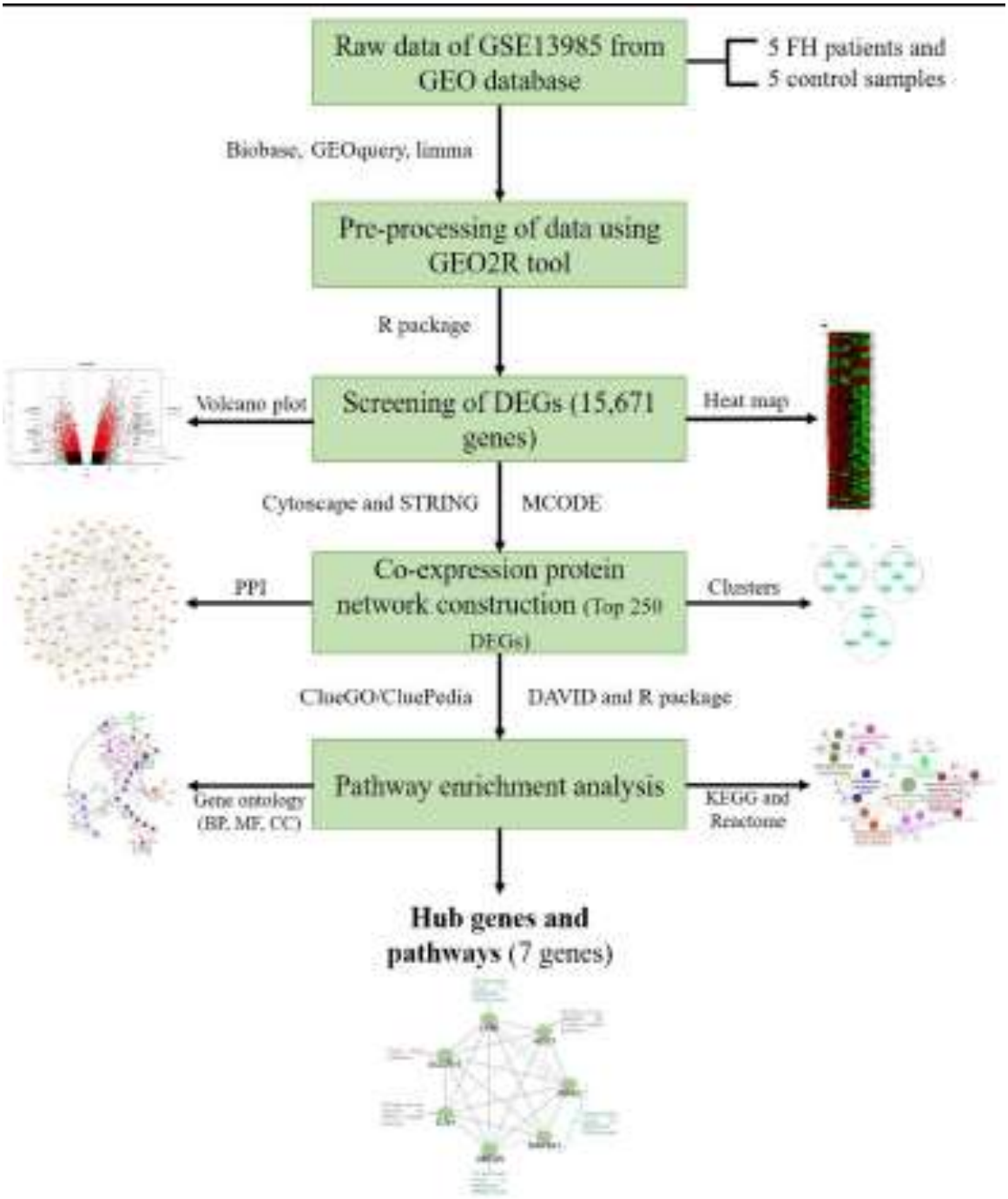
- Quantile Normalization → 1. Values are *exactly* the same between arrays (though different genes may be assigned different values in each array).
- Scale Factor Normalization → 2. Values are normally distributed with the same mean and variance across arrays.
- Invariant Set Normalization → 3. Some of the genes do not change between arrays and thus should have relatively similar values (rank invariant).

Differential gene expression analysis

- Differential expression analysis means taking the normalised read count data and performing statistical analysis to discover quantitative changes in expression levels between experimental groups.
- For example, we use statistical testing to decide whether, for a given gene, an observed difference in read counts is significant, that is, whether it is greater than what would be expected just due to natural random variation.
- **Methods for differential expression analysis**
- There are different methods for differential expression analysis such as edgeR and DESeq based on negative binomial (NB) distributions or baySeq and EBSeq which are Bayesian approaches based on a negative binomial model.
- It is important to consider the experimental design when choosing an analysis method.
- While some of the differential expression tools can only perform pair-wise comparison, others such as edgeR,



limma-voom, DESeq and maSigPro can perform multiple comparisons.



- GEO2R performs comparisons on original submitter-supplied processed data tables using the GEOquery and limma R packages from the Bioconductor project.
- Bioconductor is an open source software project based on the R programming language that provides tools for the analysis of high-throughput genomic data.

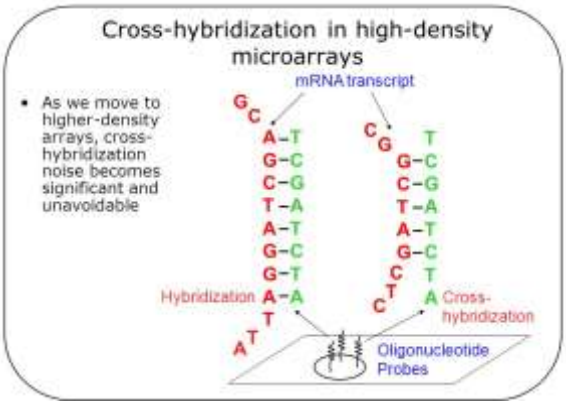
- The *GEOquery* R package parses GEO data into R data structures that can be used by other R packages. The *limma* (Linear Models for Microarray Analysis) R package has emerged as one of the most widely used statistical tests for identifying differentially expressed genes.
- It handles a wide range of experimental designs and data types and applies multiple-testing corrections on P-values to help correct for the occurrence of false positives. Thus, GEO2R provides a simple interface that allows users to perform R statistical analysis without command line expertise.

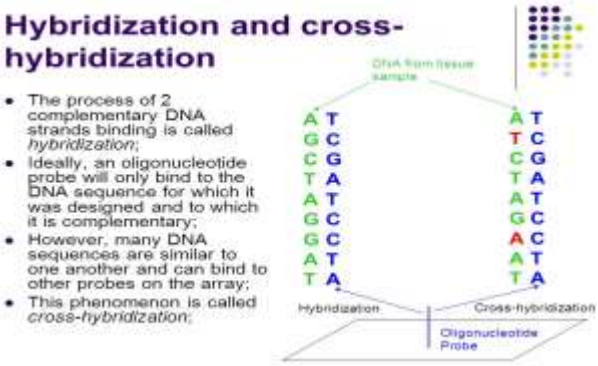
UNIT – 3- SBIA5304-MICROARRAY DATAANALYSIS

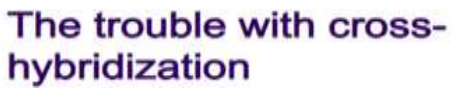
UNIT III PREDICTION

- Prediction of cross hybridization to related genes,
- Thermodynamics of nucleic acid duplexes,
- Prediction of Tm- probe
- Dimensionality reduction, principal component analysis,
- Machine learning methods for cluster analysis; Hierarchical clustering
- Analysis of relationships between genes, tissues or treatments- similarity of gene or sample profiles –Classification of tissues and samples – validation.

- Cross-hybridization is the tendency for chains of nucleic acids to bind to other chains of nucleic acids that have similar but not identical sequences.
- This has the potential to make the interpretation of microarray experiments difficult since intensity at a spot on the array does not simply depend on the quantity of target in the sample.







- With cross-hybridization, each probe will signal the presence of multiple sequences other than that it was designed for;
- This skews the observed data from the expected data.

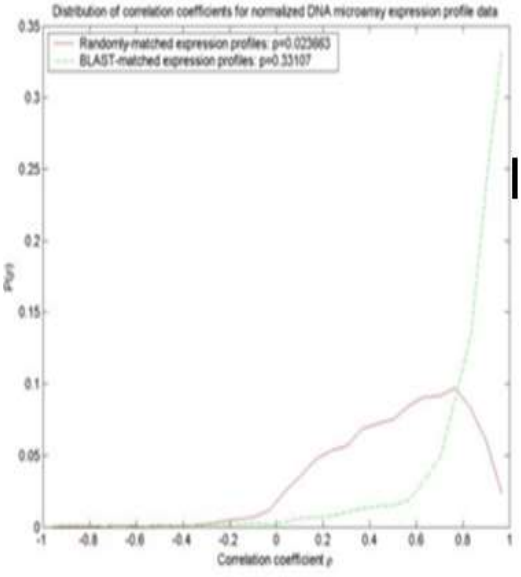
Expected expression profile vector (no hybridization)

Observed expression profile vector (cross-hybridized)

Detecting cross-hybridization

- To test for whether cross-hybridization is impacting the gene expression data, we perform a BLAST sequence match on all oligonucleotide probe sequences used on the microarray;
- Many probes will be matched with sequences for which it wasn't specifically designed.

- We compute the Pearson correlation coefficient ρ between matched probe sequence expression profiles and between the profiles of randomly-paired probes;
- Approximately 33% of the BLAST-matched probes have ρ > 0.95, whereas only 2% of randomly-matched probes have ρ >0.95;
- This difference in the 2 distributions indicates that cross-hybridization indeed has a *significant impact* on the observed gene expression data.



CrossHybDetector: detection of cross-hybridization events in DNA microarray experiments

- DNA microarrays contain thousands of different probe sequences represented on their surface.
- These are designed in such a way that potential cross-hybridization reactions with non-target sequences are minimized.
- However, given the large number of probes, the occurrence of cross hybridization events cannot be excluded.
- This problem can dramatically affect the data quality and cause false positive/false negative results.
- *CrossHybDetector* is a software package aimed at the identification of cross-hybridization events occurred during individual array hybridization, by using the probe sequences and the array intensity values.
- As output, the software provides the user with a list of array spots potentially 'corrupted' and their associated p-values calculated by Monte Carlo simulations.
- Graphical plots are also generated, which provide a visual and global overview of the quality of the microarray experiment with respect to cross-hybridization issues.

- *CrossHybDetector* is a software package aimed at the identification of cross-hybridization events occurred during individual array hybridization, by using the probe sequences and the array intensity values.
- As output, the software provides the user with a list of array spots potentially 'corrupted' and their associated p-values calculated by Monte Carlo simulations.
- Graphical plots are also generated, which provide a visual and global overview of the quality of the microarray experiment with respect to cross-hybridization issues.

R is ‘GNU S’, a freely available language and environment for statistical computing and graphics which provides a wide variety of statistical and graphical techniques: linear and nonlinear modelling, statistical tests, time series analysis, classification, clustering, etc.

CRAN is a network of ftp and web servers around the world that store identical, up-to-date, versions of code and documentation for R.

CrossHybDetector is implemented as a package within the statistical computing environment R.

Functions of *marray* and *methods R* packages are internally utilized and are required by *CrossHybDetector* to work.

Data formats

CrossHybDetector algorithm uses as input data

i) the array probe sequences, ii) the spot intensities and array layout, iii) the spot type information (i.e. for each spot, whether it is "standard probe", "negative control", "spike-in").

This information is respectively contained into three separated text files.

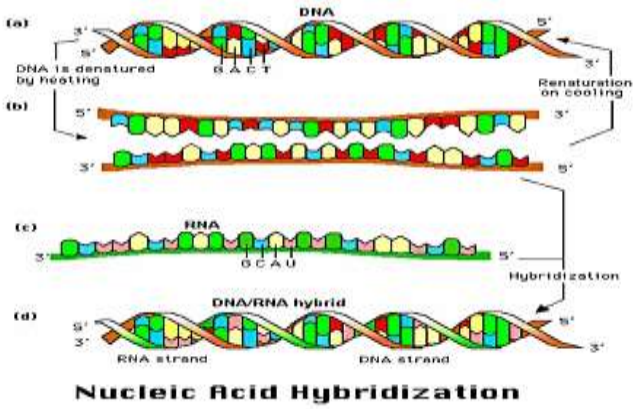
Thermodynamics of nucleic acid duplexes

- Nucleic acid thermodynamics is the study of how temperature affects the nucleic acid structure of double-stranded DNA (dsDNA).
- The melting temperature (Tm) is defined as the temperature at which half of the DNA strands are in the random coil or single-stranded (ssDNA) state.

- Tm depends on the length of the DNA molecule and its specific nucleotide sequence.
- DNA, when in a state where its two strands are dissociated (i.e., the dsDNA molecule exists as two independent strands), is referred to as having been denatured by the high temperature.

DNA hybridization

- DNA is a nucleic acid that contains the genetic instructions monitoring the biological development of all cellular forms of life, and many viruses.
- DNA is a long polymer of nucleotides and encodes the sequence of the amino-acid residues in proteins using the genetic code, a triplet code of nucleotides. DNA it is organized as two complementary strands, head-to-toe, with the hydrogen bonds between them.
- Each strand of DNA is a chain of chemical “building blocks”, called nucleotides, of which there are four types: adenine (A), cytosine (C), guanine (G) and thymine (T).
- Between the two strands, each base can only bond with one single predetermined other base: A with T, T with A, C with G, and G with C, being the only possible combination.
- Hybridization refers to the annealing of two nucleic acid strands following the base pairing rule.
- As shown in Fig at high temperatures approximately 90°C to 100°C the complementary strands of DNA separate, denature, yielding single-stranded molecules.
- Two single strands under appropriate conditions of time and temperature e.g. 65°C, will re-naturate to form the double stranded molecule.
- Nucleic acid hybrids can be formed between two strands of DNA, two strands of RNA or one strand of DNA and one of RNA.
- Nucleic acids hybridization is useful in detecting DNA or RNA sequences that are complementary to any isolated nucleic acid.



Nucleic Acid Hybridization

- The nucleic acid duplex stability can be endangered by the interaction between the nucleotide bases. Thermodynamics for double helix formation of DNA/DNA, RNA/RNA or DNA/RNA can be estimated with nearest neighbour parameters.
- Enthalpy change, ΔH∘, entropy change, ΔS∘, free energy change, ΔG∘, and melting temperature, Tm, were obtained on the basis of the nearest-neighbour model.
- The nearest-neighbour model for nucleic acids, known as the NN model, assumes that the stability of a given base pair depends on the identity and orientation of neighbouring base pairs
- In the NN model, sequence dependent stability is considered in terms of nearest-neighbour doublets.
- In duplex DNA there are 10 such unique internal nearest-neighbour doublets.
- Listed in the 5'-3' direction, these are AT/AT TA/TA AA/TT AC/GT CA/TG TC/GA CT/AG CG/CG GC/GC and GG/CC.
- Dimmer duplexes are represented with a slash separating strands in antiparallel orientation e.g. AC/TG means 5'-AC-3' Watson–Crick base-paired with 3'-TG-5'.
- The total difference in the free energy of the folded and unfolded states of a DNA duplex can be approximated at 37o, with a nearest-neighbour model:

$$\Delta G^{o}\,(\text{total}) = \sum_{i} n_i \Delta G^{o}\,(i) + \Delta G^{o}\,(\text{init w/term G} \cdot \text{C}) + \Delta G^{o}\,(\text{init w/term A} \cdot \text{T}) + \Delta G^{o}\,(\text{sym})$$

E3

where $\Delta G^{\circ}\,(i)$ are the standard free-energy changes for 10 possible Watson-Crick nearest neighbours, *e.g.* $\Delta G^{o}\,(1) = \Delta G^{o}_{37}\,(\text{AA/TT})$., $\Delta G^{o}\,(2) = \Delta G^{o}_{37}\,(\text{TA/AT})$., n_i is the number of occurrences of each nearest neighbour, i, and $\Delta G^{o}\,(\text{sym})$ equals +0.43 kcal/mol if the duplex is self complementary and zero if it is not self-complementary. The total difference in the free energy at 37°, ΔG^{o}_{37}, can be computed from ΔH^{o} and ΔS^{o} parameters using the equation:

$$\Delta G^{o}_{37} = \Delta H^{o} - T\Delta S^{o}$$

E4

For a specific temperature one can compute the total free energy using the values from Table 1. As described in [19] the melting temperature T_m is defined as the temperature at which half of the strands are in double helical and half are in the random-coil state. A random-coil state is a polymer conformation where the monomer subunits are oriented randomly while still being bonded to adjacent units.

For self-complementary oligonucleotides, the T_m for individual melting curves was calculated from the fitted parameters using the following equation:

$$T_m = \Delta H^{o} / (\Delta S^{o} + R \ln C_T)$$

E5

where R is the general gas constant, *i.e.* 1.987cal/K mol, the C_T is the total strand concentration, and T_m is given in K. For non-self-complementary molecules, C_T in equation (5) was replaced by $C_T/4$.

Sequence	ΔG 37 (kcal/mol)			
	Delcourt et al.	*SantaLucia et al.*	*Sugimoto et al.*	*Allawi et al.*
AA/TT	-0.67	-1.02	-1.20	-1.00
AT/TA	0.62	-0.73	-0.90	-0.88
TA/AT	-0.70	-0.60	-0.90	-0.58
CA/GT	-1.19	-1.38	-1.70	-1.45
GT/CA	-1.28	-1.43	-1.50	-1.44
CT/GA	-1.17	-1.16	-1.50	-1.28
GA/CT	-1.12	-1.46	-1.50	-1.30
CG/GC	-1.87	-2.09	-2.80	-2.17
GC/CG	-1.85	-2.28	-2.30	-2.24
GG/CC	-1.55	-1.77	-2.10	-1.84
Average	-1.20	-1.39	-1.64	-1.42
Init. w/term G•C	NA	0.91	1.70	0.98
Init. w/term A-T	NA	1.11	1.70	1.03

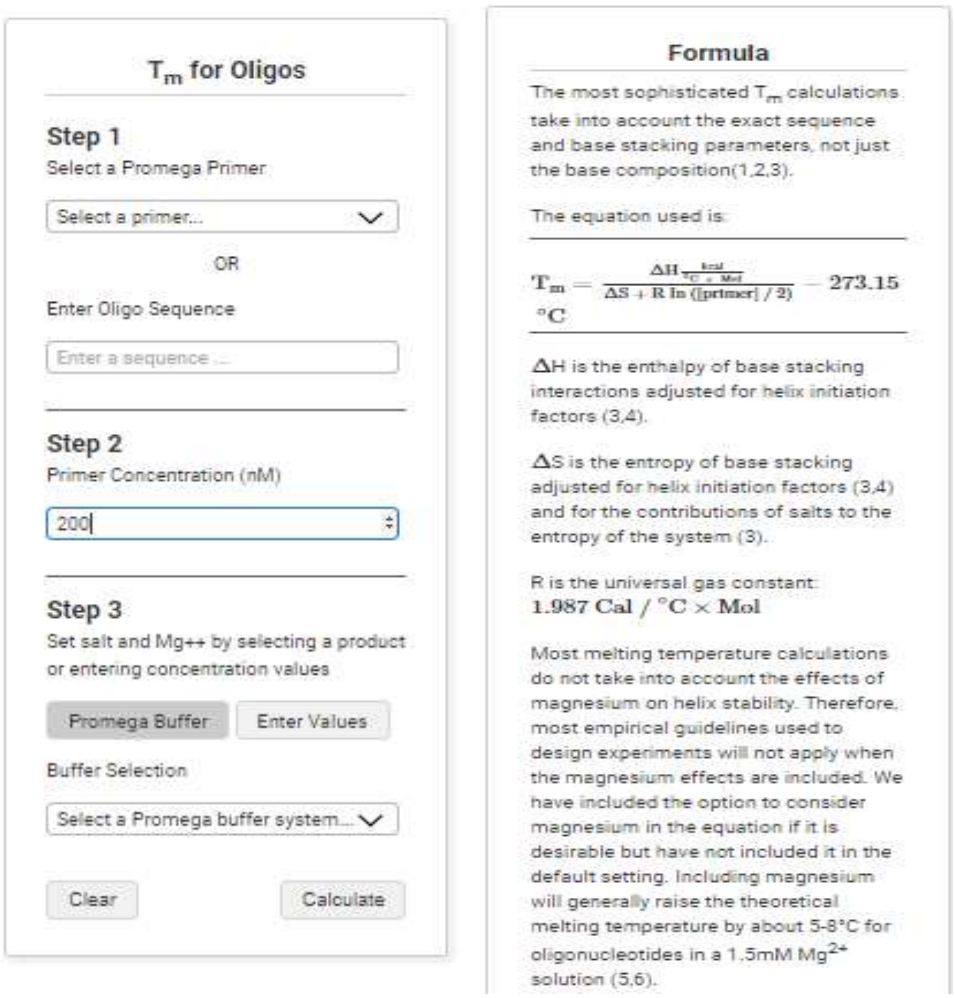
Table 2.
Comparison of computed NN free energy parameters at 37°C

Sequence	ΔH° kcal/mol	ΔS° kcal/mol
AA/TT	-7.9	-22.2
AT/TA	-7.2	-20.4
TA/AT	-7.2	-21.3
CA/GT	-8.5	-22.7
GT/CA	-8.4	-22.4
CT/GA	-7.8	-21.0
GA/CT	-8.2	-22.2
CG/GC	-10.6	-27.2
GC/CG	-9.8	-24.4
GG/CC	-8.0	-19.9
Init. w/term G•C	0.1	-2.8
Init. w/term A•T	2.3	4.1
Symmetry correction	0	-1.4

Table 1.
Unified oligonucleotide $\Delta H°$ and $\Delta S°$ nearest neighbour parameters in 1M NaCl. The table shows the values of the total enthalpy and entropy for the dimmer duplexes as used in [3].

- **The nearest-neighbour parameters** of Delcourt et al., SantaLucia et al., Sugimoto et al. and Allawi et al.were evaluated from the analysis of optical melting curves of a variety of short synthetic DNA duplexes in 1 M Na+.
- The observed trend in nearest-neighbor stabilities at 37°C is GC/CG = CG/GC > GG/CC > CA/GT = GT/CA = GA/CT = CT/GA > AA/TT > AT/TA > TA/AT, as in Table 2.
- This trend suggests that both sequence and base composition are important determinants of DNA duplex stability. It has long been recognized that DNA stability depends of the percent G-C content.

Prediction of Tm- probe



Dimensionality reduction in microarray

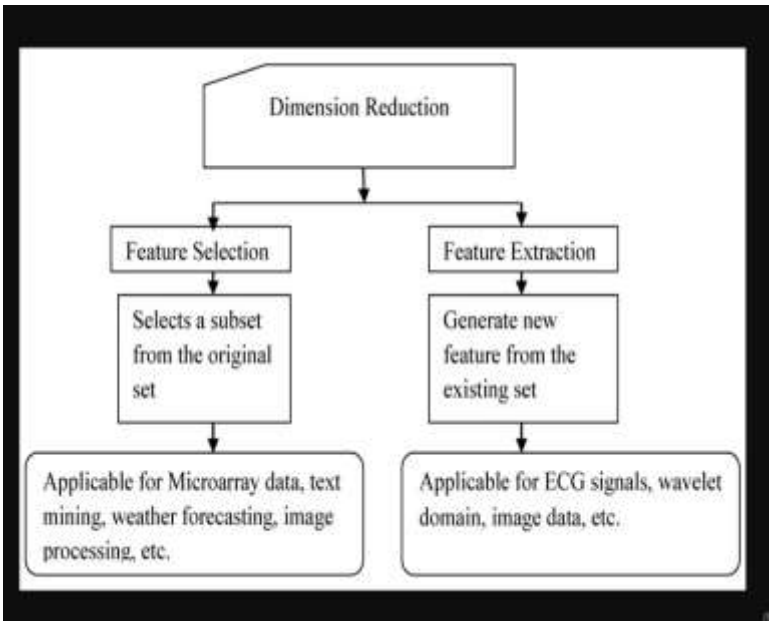
- The number of input variables or features for a dataset is referred to as its dimensionality.
- Dimensionality reduction refers to techniques that reduce the number of input variables in a dataset.
- More input features often make a predictive modeling task more challenging to model, more generally referred to as the curse of dimensionality.
- High-dimensionality statistics and dimensionality reduction techniques are often used for data visualization. Nevertheless these techniques can be used in applied machine learning to simplify a classification or regression dataset in order to better fit a predictive model.

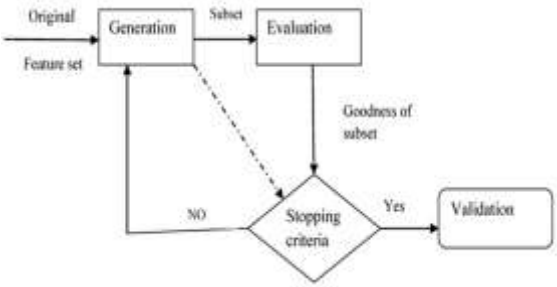
Dimensionality Reduction

Dimensionality reduction refers to techniques for reducing the number of input variables in training data.

> *When dealing with high dimensional data, it is often useful to reduce the dimensionality by projecting the data to a lower dimensional subspace which captures the "essence" of the data. This is called dimensionality reduction.*

- Traditionally manual management of the high dimensional data set is more challenging.
- With the advent of data mining and machine learning techniques, knowledge discovery and recognition of patterns from these data can be done automatically.
- However, the data in the database is filled with a high level of noise and redundancy.
- One of the reasons causing noise in these data is an imperfection in the technologies that collected the data and the source of the data itself is another reason.
- Dimensionality reduction is one of the famous techniques to remove noisy (i.e. irrelevant) and redundant features.
- For data mining techniques such as classification and clustering dimensionality reduction is treated as preprocessing task for better performance of the model.
- Dimensionality reduction techniques can be classified mainly into **feature extraction and feature selection.**
- Feature extraction approaches set features into a new feature space with lower dimensionality and the newly constructed features are usually combinations of original features.
- On the other hand, the objective of feature selection approaches is to select a subset of features that minimize redundancy and maximize relevance to the target such as the class labels in classification.
- Therefore, both feature extraction and feature selection are capable of improving learning performance, lowering computational complexity, building better-generalized models, and decreasing required storage.
- Fig shows **the classification of dimension reduction process** and the data set in which these are generally applied in the literature.

- Feature selection selects a group of features from the original feature set without any changeover and maintains the physical meanings of the original features.
- Therefore, feature selection is superior in terms of better readability and interpretability.
- One of the applications would be in gene microarray data analysis.
- Feature selection has its significance in many real-world applications such as finding relevant genes to a specific disease in Microarray data, analysis of written text, and analysis of medical images, analysis of the image for face recognition and for weather forecasting.
- 



There are four basic stages in feature selection method:

Generation Procedure (GP), to select candidate feature subset

Evaluation Procedure (EP), to evaluate the generated candidate feature subset and output, a relevancy value

Stopping Criteria (SC): To determine when to stop

Validation Procedure (VP): To determine whether it is the optimal feature subset or not.

Generation Procedure (GP)

This procedure generates a subset of features that is relevant to the target concept.

GP are of two types

Individual Ranking

Measures the relevance of each feature. The feature relevance is measured based on some evaluation function. In this case, each individual feature is evaluated by assigning some weight or score.

Subset Selection

A subset of features is selected based on some search strategy. If the size of the data set is N×M, then a total number of features in the data set is N. The possible number of subsets of features is 2^N. This is even very large for a medium sized feature set. Therefore suitable search strategy is applied to this process.

The search is classified as:

A. Complete: It traverses all the feasible solutions. This procedure does an exhaustive search for the best possible subset pertaining to the evaluation function. Example of complete search is a branch and bound best first search.

B Heuristic Deterministic: uses a greedy strategy to select features according to local change. There are many alternatives to this straightforward method, but the creation of subset is basically incremental. Examples of this procedure are sequential forward selection, sequential backward selection, sequential floating forward selection, and sequential floating backward selection.

C. Nondeterministic (Random): It attempts to find an optimal solution in a random fashion. This procedure is new in the field of feature selection methods compared to the above two categories. Optimality of the selected subset depends on the resources available.

Evaluation Procedure (EP)

An optimal subset is always relative to a certain evaluation function. An evaluation function tries to measure the discriminating ability of a feature or a subset to distinguish the different class labels.

The evaluation function is categorized as distance, information (or uncertainty), dependence, consistency, and classifier error rate.

Distance Measures

For a two-class problem say A and B are two features, then A and B are selected on the basis of their distance (e.g. Euclidian distance). If the distance is zero then the features are said to be redundant and ignored. The higher the distance the more the features are discriminating.

Information Measures

This determines the information gain for the feature. Feature A is preferred over feature B if the information gain of A is more than B (e.g. entropy measure).

Dependence Measures

Dependence or correlations of the ability to predict the value of one variable from the value of another. If the correlation of feature A with class C is higher than the correlation of feature B with class C then feature A is preferred to B. This measure finds the minimally sized subset that satisfies the acceptable inconsistency rate that is usually set by the user.

Consistency Measure

This measure finds the minimally sized subset that satisfies the acceptable inconsistency rate that is usually set by the user.

Classifier Error Rate

The evaluation function is the classifier itself. It measures the accuracy of the classifier for different subsets of feature set and measures the error rate for the different subset. We have classified the feature selection method as non-soft computing based and soft computing based. Based on the generation procedure and evaluation function, the feature selection methods are classified, where the generation procedure and evaluation functions are two dimensions.

Stopping Criteria

It indicates the end of the process. Commonly used stopping criteria are:

(i) When the search completes

(ii) When some given bound (minimum number of features or a maximum number of iterations) is reached.

(iii) When a subsequent addition (or deletion) of any feature does not produce a better subset and

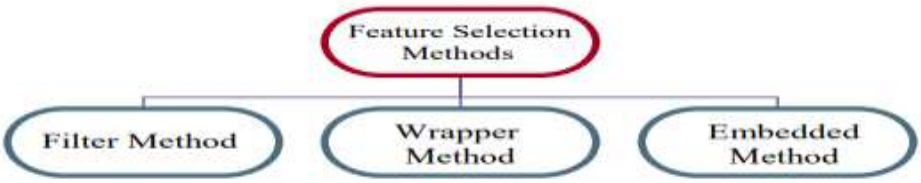
(iv) When a sufficiently good subset (e.g. a subset if its classification error rate is less than the allowable error rate for a given task) is selected.

Feature selection approaches are primarily categorized as a **filter, wrapper, and embedded method.**

Recently other feature selection methods are gaining popularity i.e., hybrid and ensemble methods (Fig).

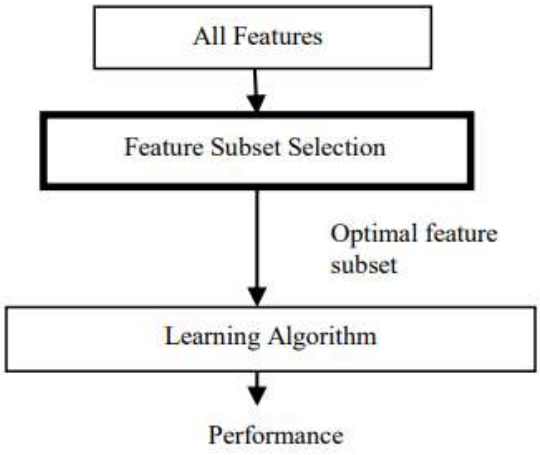
Table 1. Classification of feature selection methods based on combination of GP and EF.

<table>
<tr><th rowspan="2">Generation Procedure (GP)</th><th colspan="5">Evaluation Function(EF)</th></tr>
<tr><th>Distance</th><th>Information</th><th>Correlation</th><th>Consistency</th><th>Classifier error rate</th></tr>
<tr><td>Heuristic</td><td colspan="4" rowspan="3">Filter approach</td><td rowspan="3">Wrapper approach</td></tr>
<tr><td>Complete</td></tr>
<tr><td>Random</td></tr>
<tr><td colspan="6">Embedded approach (filter + wrapper)</td></tr>
</table>

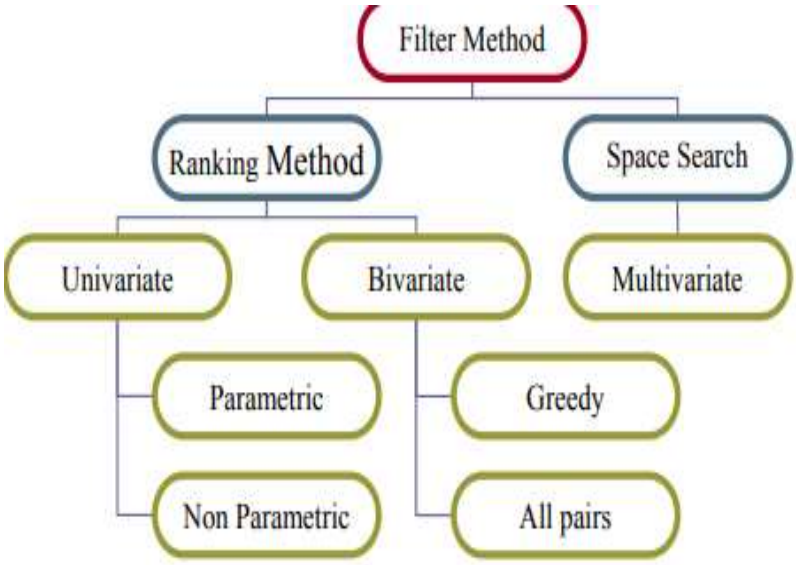


Classification of feature selection methods.

- Filter method deals with individual ranking as well as subset selection.
- The individual ranking is based on the evaluation functions such as distance, information, dependence, and consistency excluding the classifier (Fig).
- Filter techniques judge the relevance of genes by looking only at the intrinsic properties of the data. In microarray data, a gene relevance score is calculated, and low-scoring genes are removed. Afterward, this subset of genes is presented as input to the classification algorithm.
- The filtering technique can be used as a pre-processing step to reduce space dimensionality and overcome overfitting.



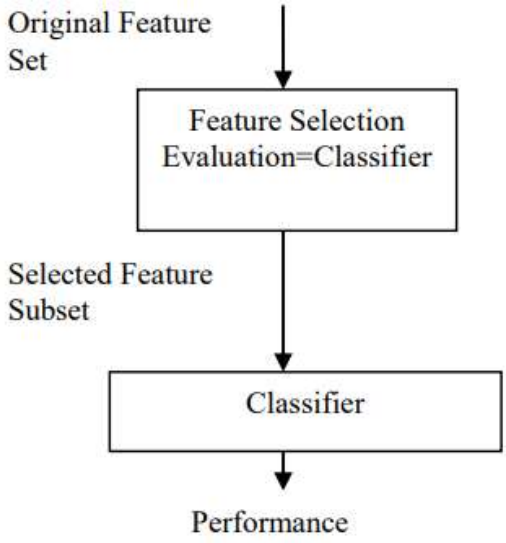
- The filter approach is commonly divided into two different sub-classes:
- **Individual evaluation and subset evaluation.**
- In individual evaluation method, the gene expression dataset is given as input. The method has an inbuilt evaluation process according to which a rank is provided to each individual gene based on which the selection is done.
- Different criteria can be adopted, like setting a threshold for the scores and selecting the genes which satisfy the threshold criteria, or sometimes the threshold can be chosen in such a way that a maximum number of genes can be selected.
- Then, the subset selected can be the final subset which is used as the input to the classifiers. In subset selection, all GP and evaluation function excluding the classifier can be taken for the combination
- However, methods in this framework may suffer from an inevitable problem, which is caused by searching through
- the possible feature subsets.
- The subset generation process usually increases the computational time but gives more relevant feature subset.
- In literature, it is found that the subset evaluation approach outperformed the ranking methods
- The filter method is again classified into the ranking method and space search method.
- Fig.describes the taxonomy of filter feature selection method.



- Taxonomy of filter FS methods: Pros of Filter Feature Selection Method.
- The method is simple and fast.
- It scales well to high dimensional data.
- It is independent of classifiers.
- Cons of Filter Feature Selection Method
- The method is generally univariate or low variate.

Wrapper Method

- In the wrapper approach, all GP can be taken in combination with the classifier as evaluation function and generates the relevant feature subset.
- Wrappers are feedback methods, which incorporate the machine-learning algorithm in the feature selection process, i.e., they rely on the performance of a specific classifier to evaluate the quality of a set of features.
- **Wrapper methods search through the space of feature subsets** and calculate the estimated accuracy of a single learning algorithm for each feature that can be added to or removed from the feature subset.
- The search may be a GP and the **evaluation function is a classifier.**



Wrapper Method

1. How to find all possible feature subsets for evaluation?
2. How to satisfy oneself with the classification performance of the chosen classifier in order to guide the search and what should be the stopping criteria?

Which predictor to use?

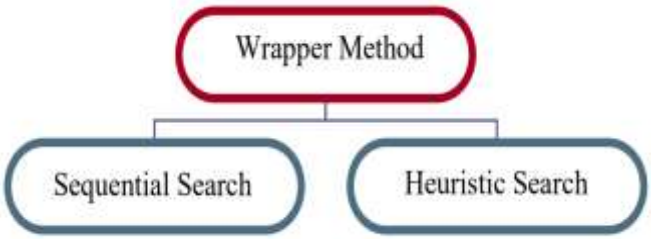
The wrapper approach applies a blind search to find a subset of features. It searches randomly for the best subset which cannot be made sure without getting all possible subsets. Therefore, feature selection in this approach is NP-hard and the search with each iteration tends to become intractable for the user. This is not a preferred approach for feature selection, as it is a crude force method and requires higher computational time for feature subset selection.

The feature space in case of wrapper approach can be searched with various strategies, *e.g.*, forward (*i.e.*, by adding attributes to an initially empty set of attributes) or backward (i e., by starting with the full set and deleting attributes one at a time). The correctness of a specific subset of features/genes based on our classifier is obtained by training and testing the subset against that specific classification model.

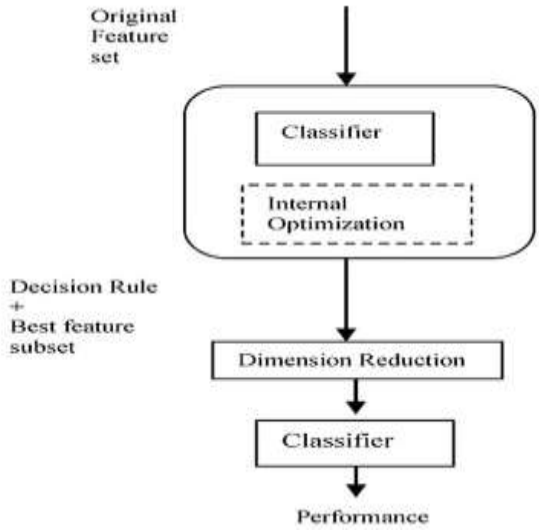
The advantage of wrapper approach is that it selects a near perfect subset and error rate in this method is less as compared to other methods. The major disadvantage of the method is that it is computationally very intensive and it is intended for the particular learning machine on which it has been tested. Therefore, there is a high risk of overfitting than filter techniques.

The wrapper approach of feature selection is classified as sequential search based and Heuristic search based.

The taxonomy of the wrapper model is given in Fig

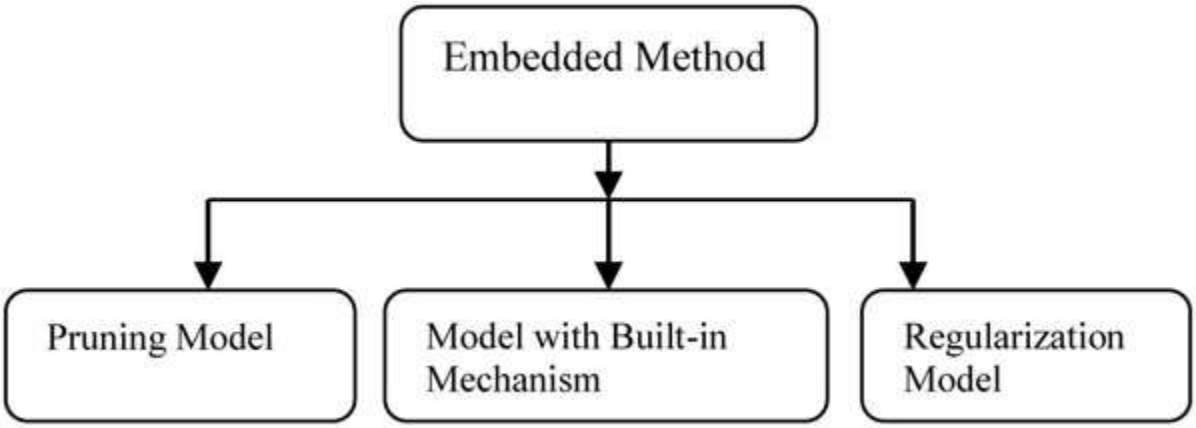


- Usually, an exhaustive search is too expensive, and thus non-exhaustive, heuristic search techniques like genetic algorithms, greedy stepwise, best first or random search are often used.
- Here, feature selection occurs externally to the induction method using the method as a subroutine rather than as a post-processor.
- In this process, the induction algorithm is called for every subset of feature consequently inducing high computational cost
- **Embedded Method**
- Despite the lower time consumption of the filter method, a major limitation of the filter approach is that it is independent of the classifier, usually resulting in worse performance than the wrappers.
- However, the wrapper model comes with an expensive computational cost, which is particularly aggravated by the high dimensionality of microarray data.
- An intermediate solution for researchers is the use of hybrid or embedded methods, which use the core of the classifier to establish criteria to rank features.
- Embedded methods are more tractable and efficient in comparison to wrapper approach.
- This method has a lower risk of overfitting compared to wrapper approach. Probably the most famous embedded method is Support Vector Machine based on Recursive Feature Elimination (SVM-RFE).



The embedded method is classified into three different categories.

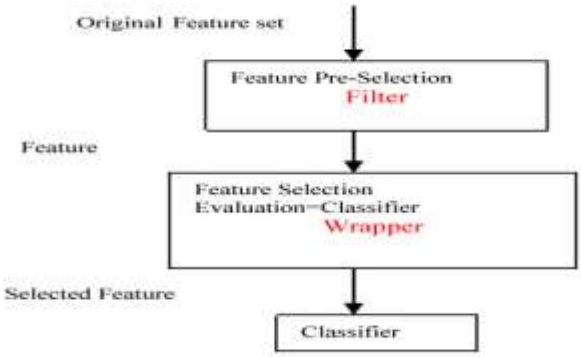
The taxonomy of embedded method is shown in Fig.

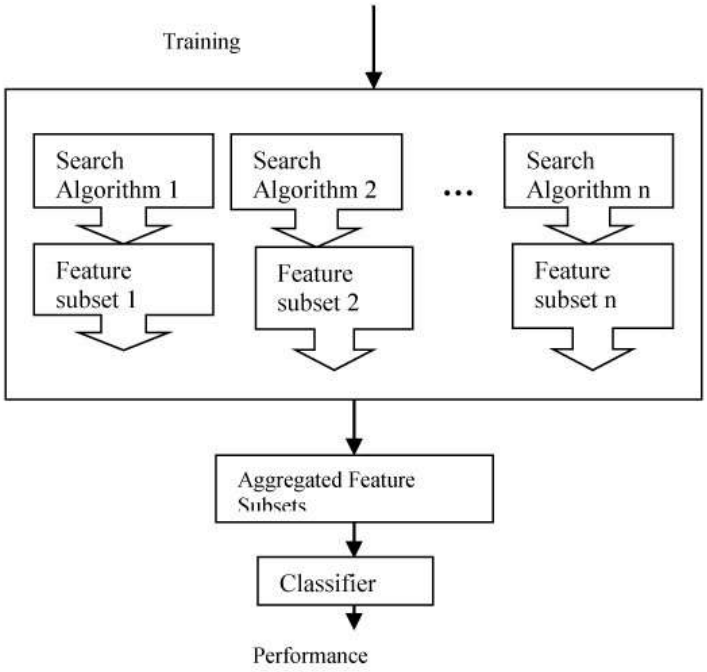


Hybrid Method

- It is the combination of any number of same or different classical methods of feature selection such as filter and wrapper methods.
- The combination can be a filter-filter, filter-wrapper, and filter-filter-wrapper where the gene subset obtained from one method is served as the input to another selection algorithm.
- Generally, filter is used to select the initial gene subset or help to remove redundant genes.

- Any combination of several filter techniques can be applied vertically to select the preliminary feature subset.
- In the next phase, the selected features are given to the wrapper method for the optimal feature selection. This method uses different evaluation criteria.
- Therefore, it manages to improve the efficiency and prediction accuracy with the better computational cost for high dimensional data.





Clustering Techniques Analysis for Microarray Data

- There are two major types of microarray experiments: cDNA microarray and oligonucleotide arrays .

- Both the experiments consist of basic three steps: first is chip manufacturing, second are target preparation, labelling and hybridization and third is the scanning process.
- Gene expression data is expressed in form of expression matrix having real values showing the protein level of a particular gene.
- Gene expression data contains thousands of genes but less number of samples.
- There are various problems with microarray data such as:

(a) Microarray data is high dimensional data characterized by thousands of genes for small sample size, which grounds significant problems such as irrelevant and noise genes, complexity in constructing classifiers, and multiple gene-expression values are missing due to inappropriate scanning.

(b) Another drawback is mislabeled sample data or doubtful sample results by experts.

(c) Biological relevancy result is another integral criterion that should be taken into account in analyzing microarray data rather than only focusing on accuracy of cancer classification.

Clustering techniques:

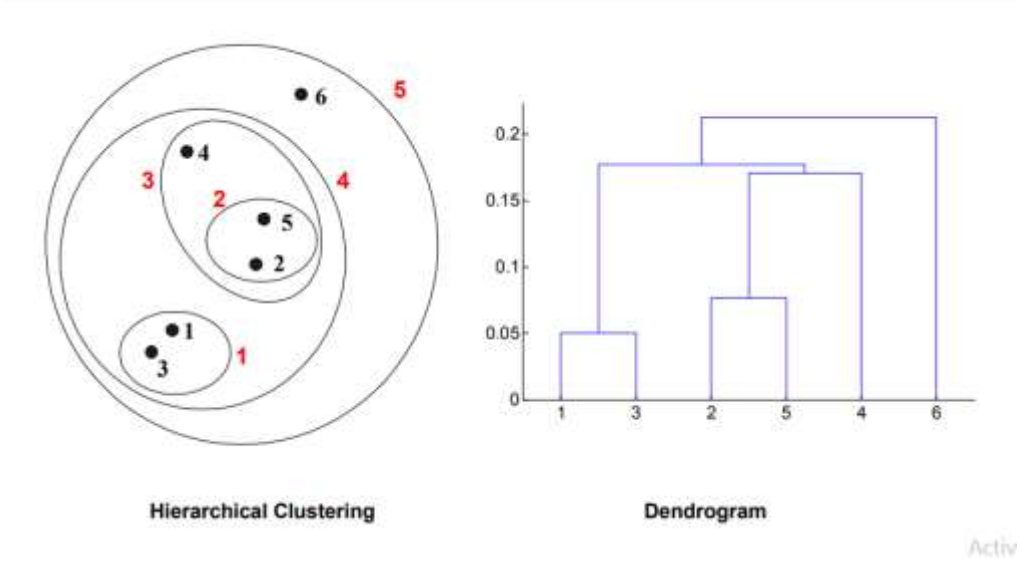
- In gene expression data, it is worth to cluster both genes and samples. There are three types of clustering that can be applied on microarray data: gene based clustering, sample based clustering and subspace clustering where genes and samples are treated in same manner.
- In case of gene clustering, the clustering is used to reduce the search dimension of the dataset.
- In case of sample based clustering, the clustering is used to group the samples of same kind whereas in subspace based clustering both the tasks are performed.
- Gene based clustering can be applied on the supervised dataset where the samples are already classified.
- The distinctive characteristic of gene expression data allows clustering both gene and samples. The clustering analysis of sampled data is to find new biological classes or to refine the existing ones

Hierarchical Clustering:

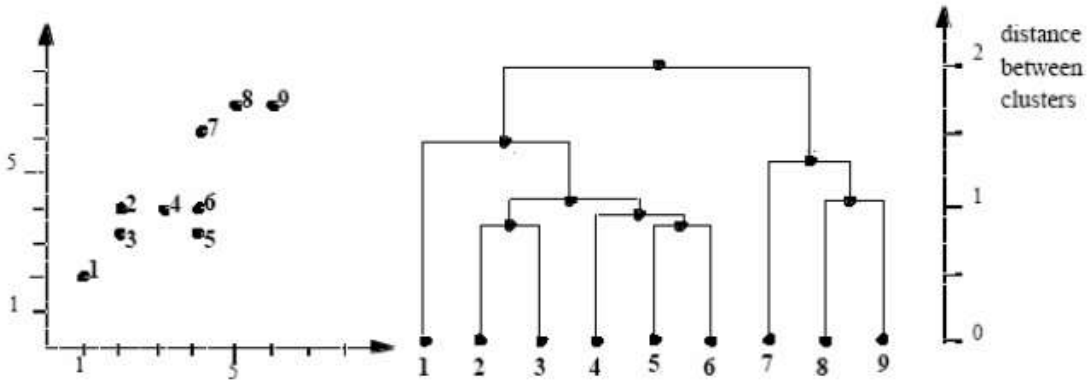
(a) **Agglomerative hierarchical clustering** –In this each object initially represents a cluster of its own. Then clusters are recursively merged until the desired cluster formation is obtained.

(b) **Divisive hierarchical clustering** - All objects initially belong to one cluster. Then the cluster is divided into sub-clusters which are successively divided into sub clusters. This process continues until the desired cluster structure is obtained.

Some commonly used metrics for hierarchical clustering are: Euclidean distance, Squared Euclidean distance, Manhattan distance, Maximum distance, Mahalanobis distance and cosine similarity



Hierarchical Clustering Dendrogram



- The *root* represents the whole data set
- A *leaf* represents a single object in the data set
- An *internal node* represent the union of all objects in its sub-tree
- The *height* of an internal node represents the distance between its two child nodes

Two main types of hierarchical clustering.

- **Agglomerative**:
  - Start with the points as individual clusters
  - At each step, merge the closest pair of clusters.
  - Until only one cluster (or k clusters) left
  - This requires defining the notion of cluster proximity.
- **Divisive:**
  - Start with one, all-inclusive cluster
  - At each step, split a cluster
  - Until each cluster contains a point (or there are k clusters)
  - Need to decide which cluster to split at each step.

Basic Agglomerative Hierarchical Clustering Algorithm

1. Initially, each object forms its own cluster
2. Compute all pairwise distances between the initial clusters (objects)

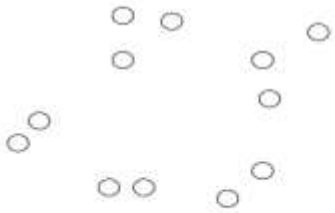
repeat

3. Merge the closest pair (A, B) in the set of the current clusters into a new cluster $C = A \cup B$
4. Remove A and B from the set of current clusters; insert C into the set of current clusters
5. Determine the distance between the new cluster C and all other clusters in the set of current clusters

until only a single cluster remains

Agglomerative Hierarchical Clustering: Starting Situation

- For agglomerative hierarchical clustering we start with clusters of individual points and a proximity matrix.

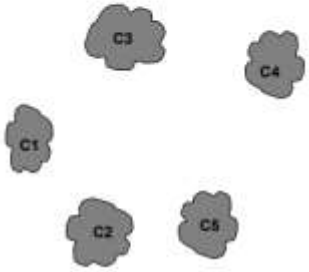


	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						
.						
.						

Proximity Matrix

Agglomerative Hierarchical Clustering: Intermediate Situation

- After some merging steps, we have some clusters.

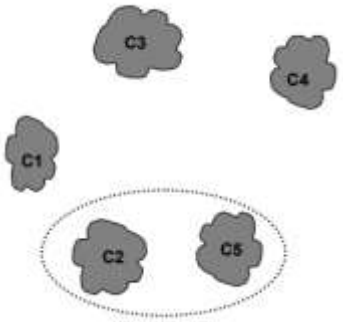


	C1	C2	C3	C4	C5
C1					
C2					
C3					
C4					
C5					

Proximity Matrix

Agglomerative Hierarchical Clustering: Intermediate Situation

- We want to merge the two closest clusters (C2 and C5) and update the proximity matrix.



	C1	C2	C3	C4	C5
C1					
C2					
C3					
C4					
C5					

Proximity Matrix

Inter-cluster distances

Four widely used ways of defining the **inter-cluster distance**, i.e., the distance between two separate clusters C_i and C_j, are

- **single linkage method** (nearest neighbor):
  $$d(C_i, C_j) = \min_{x \in C_i, y \in C_j} \{ d(x, y) \}$$
- **complete linkage method** (furthest neighbor):
  $$d(C_i, C_j) = \max_{x \in C_i, y \in C_j} \{ d(x, y) \}$$
- **average linkage method** (unweighted pair-group average):
  $$d(C_i, C_j) = avg_{x \in C_i, y \in C_j} \{ d(x, y) \}$$
- **centroid linkage method** (distance between cluster centroids c_i and c_j):
  $$d(C_i, C_j) = d(c_i, c_j)$$

Single linkage (minimum distance) method

- Distance (dissimilarity) of two clusters is based on the two most similar (closest) points in the different clusters C_i and C_j :

$$d(C_i, C_j) = \min_{x \in C_i, y \in C_j} \{ d(x, y) \}$$

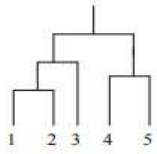
–Determined by one pair of points, i.e., by one link in the proximity graph.

–Can handle non-elliptical shapes.

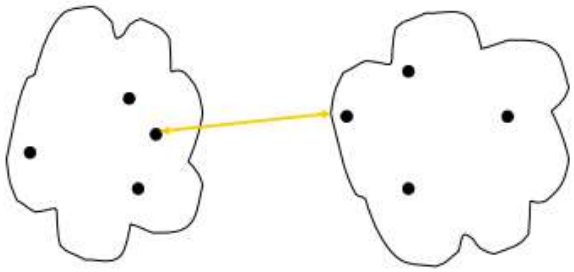
–Sensitive to noise and outliers.

Similarity matrix

	I1	I2	I3	I4	I5
I1	1.00	0.90	0.10	0.65	0.20
I2	0.90	1.00	0.70	0.60	0.50
I3	0.10	0.70	1.00	0.40	0.30
I4	0.65	0.60	0.40	1.00	0.80
I5	0.20	0.50	0.30	0.80	1.00

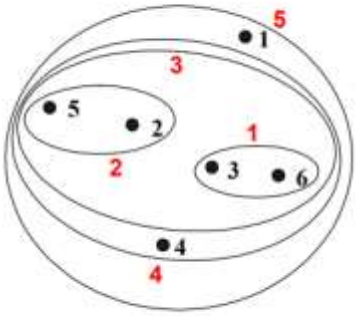


Single linkage

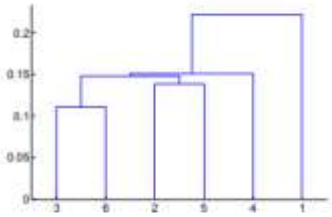


$$d(C_i, C_j) = \min_{x \in Ci, y \in Cj} \{ d(x, y) \}$$

Hierarchical Clustering: minimum distance



Nested Clusters



Dendrogram

Complete Linkage (maximum distance) method

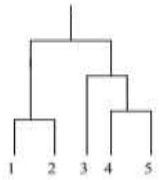
- Distance of two clusters is based on the two least similar (most distant) points in the different clusters C_i and C_j:

$$d(C_i, C_j) = \max_{x \in C_i, y \in C_j} \{ d(x, y) \}$$

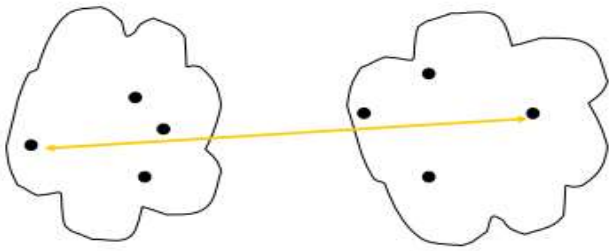
  - –Determined by all pairs of points in the two clusters.
  - –Tends to break large clusters.
  - –Less susceptible to noise and outliers.

Similarity matrix

	I1	I2	I3	I4	I5
I1	1.00	0.90	0.10	0.65	0.20
I2	0.90	1.00	0.70	0.60	0.50
I3	0.10	0.70	1.00	0.40	0.30
I4	0.65	0.60	0.40	1.00	0.80
I5	0.20	0.50	0.30	0.80	1.00



Complete linkage

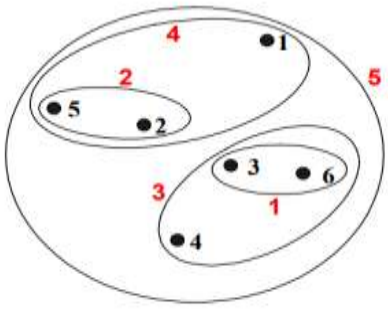


$$d(C_i, C_j) = \max_{x \in C_i, y \in C_j} \{ d(x, y) \}$$

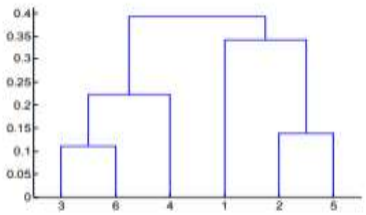
Cluster Similarity: maximum distance or Complete Linkage

- Similarity of two clusters is based on the two most distant points in the different clusters.
- Tends to break large clusters.
- Less susceptible to noise and outliers.
- Biased towards globular clusters.

Hierarchical Clustering: maximum distance

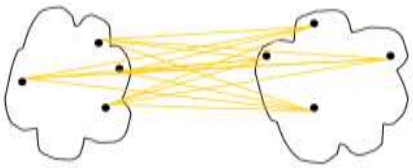


Nested Clusters



Dendrogram

Average linkage



$$d(C_i, C_j) = \frac{1}{|C_i| |C_j|} \sum_{x \in C_i} \sum_{y \in C_j} d(x, y)$$

Average linkage (average distance) method

- Distance of two clusters is the average of pairwise distances between points in the two clusters C_i and C_j:

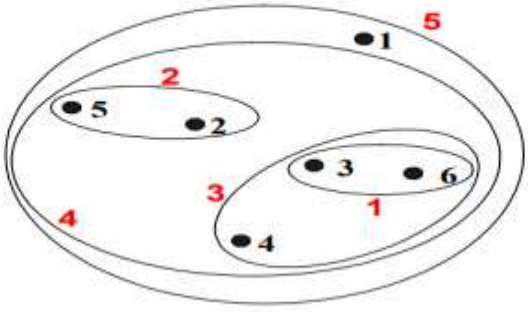
$$d(C_i, C_j) = \frac{1}{|C_i| |C_j|} \sum_{x \in C_i} \sum_{y \in C_j} d(x, y)$$

- Compromise between Single and Complete Link.
- Need to use average connectivity for scalability since total connectivity favors large clusters.
- Less susceptible to noise and outliers.
- Biased towards globular clusters.

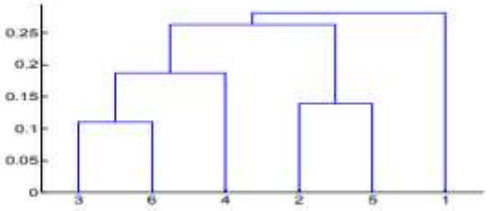
Similarity matrix

	I1	I2	I3	I4	I5
I1	1.00	0.90	0.10	0.65	0.20
I2	0.90	1.00	0.70	0.60	0.50
I3	0.10	0.70	1.00	0.40	0.30
I4	0.65	0.60	0.40	1.00	0.80
I5	0.20	0.50	0.30	0.80	1.00

Hierarchical Clustering: Average distance



Nested Clusters



Dendrogram

Centroid linkage (centroid distance) method

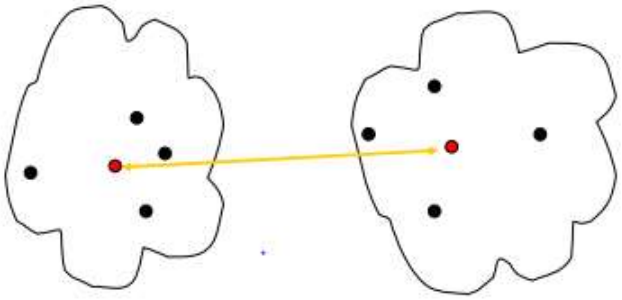
- Distance of two clusters is distance of the two centroids c_i and c_j of the two clusters C_i and C_j:

$$d(C_i, C_j) = d(c_i, c_j)$$

$$c_i = \frac{1}{|C_i|}\sum_{x \in C_i} x \qquad c_j = \frac{1}{|C_j|}\sum_{x \in C_j} x$$

- Compromise between Single and Complete Link.
- Less computationally intensive with respect to average linkage.

Centroid linkage



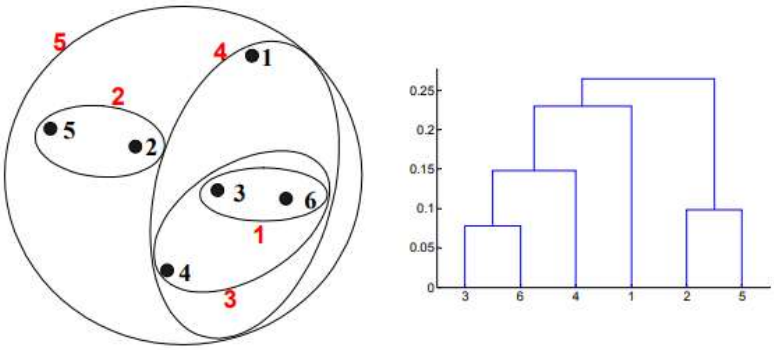
$$d(C_i, C_j) = d(c_i, c_j)$$

$$c_i = \frac{1}{|C_i|}\sum_{x \in C_i} x \qquad c_j = \frac{1}{|C_j|}\sum_{x \in C_j} x$$

Cluster Similarity: Ward's Method

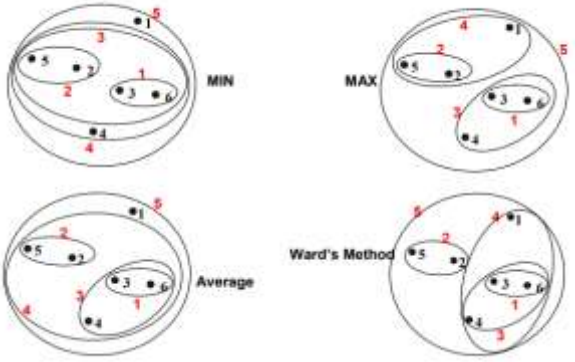
- Similarity of two clusters is based on the increase in squared error when two clusters are merged.
  - Similar to group average if distance between points is distance squared.
- Less susceptible to noise and outliers.
- Biased towards globular clusters.
- Hierarchical analogue of K-means
  - But Ward's method does not correspond to a local minimum
  - Can be used to initialize K-means

Hierarchical Clustering: Ward's method



Nested Clusters **Dendrogram**

Hierarchical Clustering: comparison



Comparison of minimum, maximum, average and centroid distance

Minimum distance

- When d min is used to measure distance between clusters, the algorithm is called the nearest-neighbor or single- linkage clustering algorithm
- If the algorithm is allowed to run until only one cluster remains, the result is a minimum spanning tree (MST)
- This algorithm favors elongated classes

Maximum distance

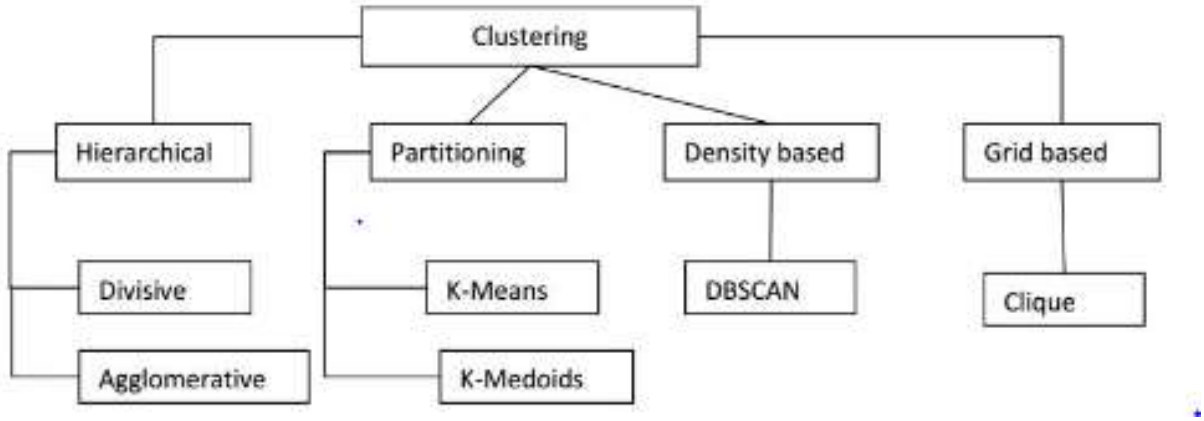
- When d max is used to measure distance between clusters, the algorithm is called the farthest-neighbor or complete- linkage clustering algorithm
- From a graph- theoretic point of view, each cluster constitutes a complete sub- graph
- This algorithm favors compact classes

Average and centroid distance

- The minimum and maximum distance are extremely sensitive to outliers since their measurement of between- cluster distance involves minima or maxima
- The average and centroid distance approaches are more robust to outliers
- Of the two, the centroid distance is computationally more attractive
- Notice that the average distance approach involves the computation of $|C_i||C_j|$ distances for each pair of clusters

Advantages and disadvantages of Hierarchical clustering

Advantages	Disadvantages
• Does not require the number of clusters to be known in advance • No input parameters (besides the choice of the (dis)similarity) • Computes a complete hierarchy of clusters • Good result visualizations integrated into the methods	• May not scale well: runtime for the standard methods: $O(n^2 \log n)$ • No explicit clusters: a "flat" partition can be derived afterwards (e.g. via a cut through the dendrogram or termination condition in the construction) • No automatic discovering of "optimal clusters"



Partitioning Algorithms:

They are iterative relocation algorithm.

They are non hierarchical or flat methods.

This method divides the data objects into non overlapping clusters such that each data object is in exactly one subset.

There are several methods which are used to implement partitioning clustering such as**:**

(a) **K-medoids,**

(b) **K-means,**

(c) **Probabilistic**

(d) K-means is one of the simplest unsupervised learning algorithms that solve the well known clustering problem. It performs the division of objects into clusters which are similar between them and dissimilar to the objects belonging to another cluster.

(e) The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters) fixed a priority. The main idea is to define k centroids, one for each cluster. These centroids should be placed in a cunning way because of different location causes different result.

(f) So, the better choice is to place them as much as possible far away from each other. The next step is to take each point belonging to a given data set and associate it to the nearest centroid.

(g) When no point is pending, **the first step is completed and an early groupage is done**. At this point we need to re-calculate k new centroids as barycenters of the clusters resulting from the previous step.

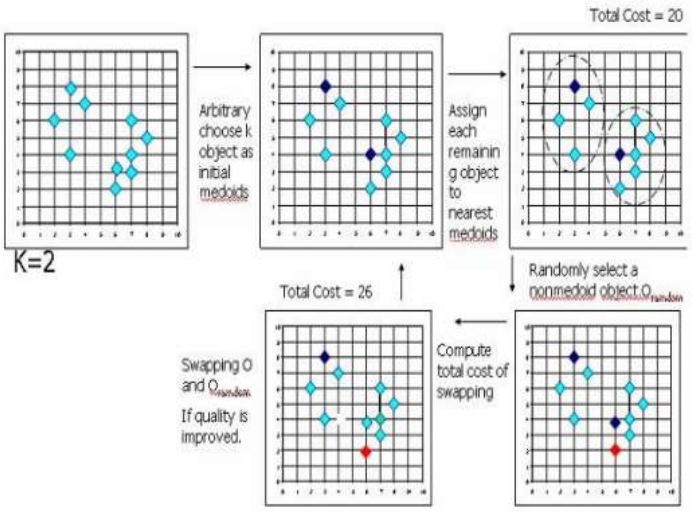
(h) After we have these k new centroids, a new binding has to be done between the same data set points and the nearest new centroid. A loop has been generated. As a result of this loop we may notice that the k centroids change their location step by step until no more changes are done.

The Euclidean distance between an object and all the nearby centroid is calculated as per the formula given below

$$j = \sum_{j=1}^{k} \sum_{i=1}^{n} \left\| x_i^{(j)} - c_j \right\|^2$$

Where $||Xi^{(j)} - Cj||^2$ is the nearest distance measure between a data point xij and the Centroid Cj, and it indicates the distance between data points from their Centroid. The time complexity of the K-means algorithm is subjected to the formula; $O(n^{dk+1})$.

Working of K-medoid Algorithm



K-Medoids Method:

- It is one of the important method of partitioning. K-medoid is based on medoids calculating by minimizing the absolute
- distance between the points and the selected centroid, rather than minimizing the square distance. As a result, it's more robust to noise and outliers than k-means. In k-medoids clustering, each cluster is represented by one of the data point in the cluster.
- These points are named cluster medoids. Here, k data objects are selected randomly as medoids to represent k cluster and remaining all data objects are placed in a cluster having medoid nearest (or most similar) to that data object. After processing all data objects, new medoid is determined which can represent cluster in a better way and the entire process is repeated.

- Again all data objects are bound to the clusters based on the new medoids. In each iteration, medoids change their location step by step. This process is continued until no any medoid move. As a result, k clusters are found representing a set of n data objects

The Distance is calculated as per formula given below (figure):

$$j = \sum_{i=1}^{k} \sum_{p \Omega c_j}^{n} \left\| P - O_j \right\|$$

The time complexity for the K-medoids algorithm is subjected to the formula; $O(k(n-2)^2)$. The efficiency and performance of the results in the cluster are directly dependent on clustering centre chosen. Hence all efforts to improve this algorithm depend on the which k cluster points are chosen as reference.

. Density based clustering:

The clusters in this are dense regions of objects in space that are separated by low density regions where cluster density is defined as each point must have a minimum number of points in its neighborhood.

(i) Based on density based connectivity e.g. DBSCAN

(ii) Based on density distribution functions e.g. DENCLUE

4. Constraint based clustering:

- Constraints are strong background information that should be satisfied.
- Constraints also reduce the search space and all the data in dataset has common property.
- e.g. in gene expression data set we have a constraint of low and high expressed genes.

5. Evolutionary Clustering:

- It is used to process time stamped data to produce a series of clustering.
- The similarity among existing data points varies along with time. Present clusters mainly depend on the current data features.
- Data is likely to change not too rapidly.
- Evolutionary clustering is useful for the following reasons: (i) consistency, (ii) noise removal (iii) smoothing (iv) cluster correspondence.

Mostly used for online document clustering

Graph Partitioning based Algorithms:

It depends on finding the minimum cut or minimum cliques in the proximity graph

Many other graph partitioning algorithms depends on eigen vectors and eigen values also.

It consists of three steps:

(i) preprocessing i.e. to covert data into graph and finding similarity between the nodes.

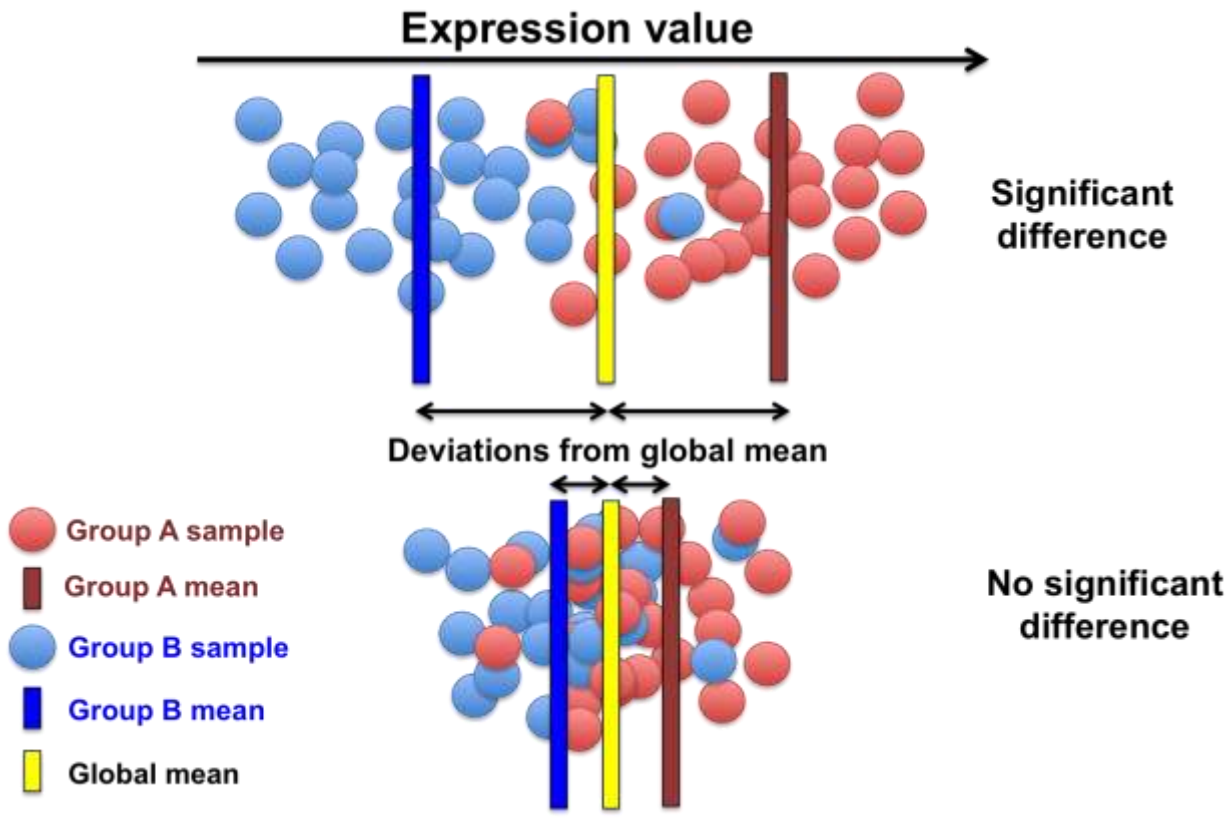
(ii) partitioning of the graph.

(iii) performing clustering until required number of clusters are not obtained.

Each clustering algorithms belongs to one of the clustering types listed above.

So that, Partitioning method is exclusive clustering, Fuzzy C-means is an overlapping clustering algorithm, Hierarchical clustering is obvious and lastly Mixture of Gaussian is a probabilistic clustering algorithm.

- Once gene expression data is obtained, one typically wishes to compare one experimental group versus a second one (or more) in order to find out which genes/transcripts change significantly between conditions.
- The process is called differential expression analysis.
- The goal of differential expression analysis is to perform statistical analysis to discover changes in expression levels of defined features (genes, transcripts, exons) between experimental groups with replicated samples.
- Essentially, it aims at comparing the **average expression of a gene in group A** with the **average expression of this gene in group B.**



Many tools exist that will perform differential expression analysis. The output of such tools is similar, and essentially revolves around interpreting:

- **Fold change**:
  For a given comparison, a positive fold change value indicates an increase of expression, while a negative fold change indicates a decrease in expression.
  This value is typically reported in **logarithmic scale (base 2)**. For example, log2 fold change of 1.5 for a specific gene in the "WT vs KO comparison" means that the expression of that gene is increased in WT relative to KO by a multiplicative factor of 2^1.5 ≈ 2.82.

- **P-value**: Indicates whether the gene analysed is likely to be differentially expressed in that comparison.
  This applies to each gene individually, **assuming that the gene was tested on its own without consideration that all other genes were also tested**. *More on P-value will follow!*

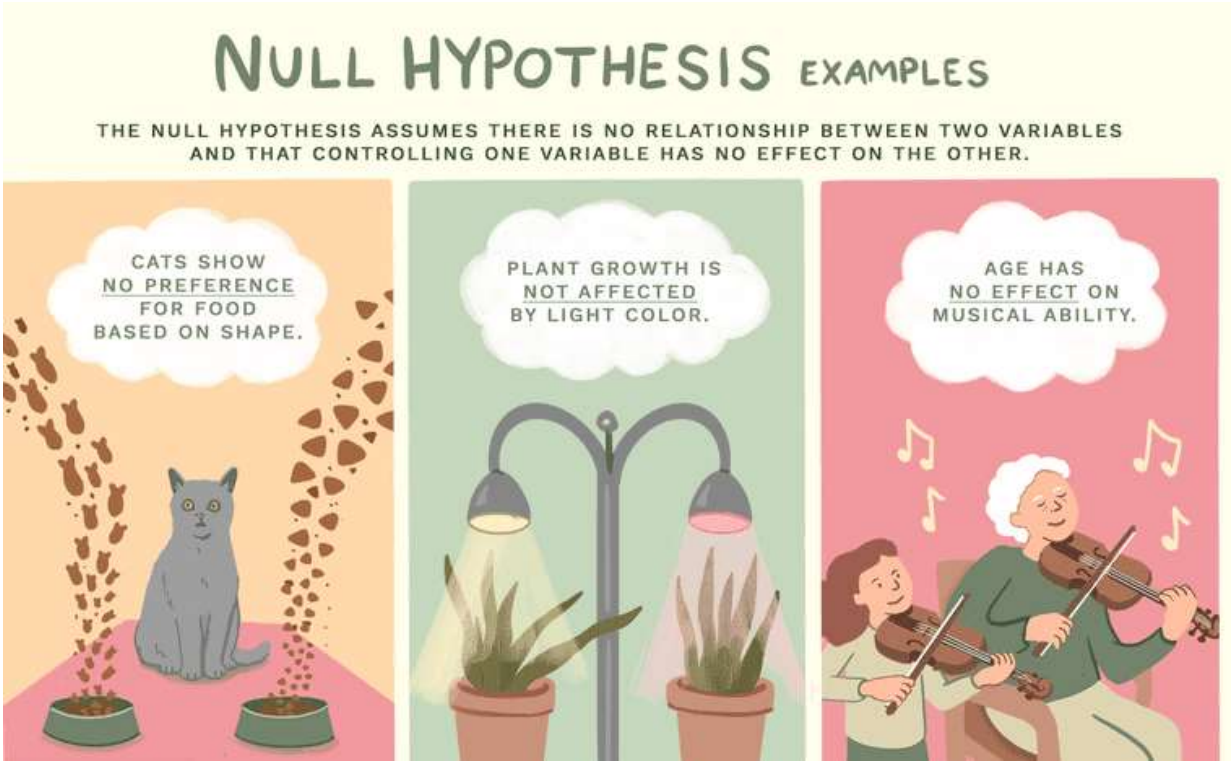
- **Adjusted (or, corrected for multiple genes testing) p-value**: The p-value obtained for each gene above is re-calculated to correct for running many statistical tests (as many as the number of genes). In the result, we can say that all genes with adjusted p-value < 0.05 are **significantly differentially expressed** in these two samples.

How to interpret P-value ?

First, we need to talk about **statistical hypothesis testing**.

Two hypotheses should be described upon designing an experiment.

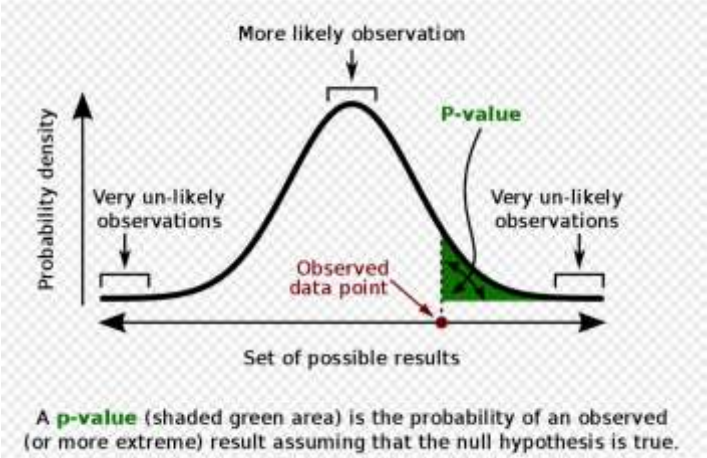
- **The NULL hypothesis *H0***: that there is NO difference, for example, of means of weight between two populations of subjects.
- **The alternative hypothesis *H1***: there is difference.



- Once the hypothesis are drawn and the significance level set, we perform a statistical test… and we obtain a p-value.
- If the p-value is below the significance level (for example, 0.05), we can reject the null hypothesis in favor of the alternative hypothesis, i.e. we conclude that the observed difference is the result of a real effect.
- Once the hypothesis are drawn and the significance level set, we perform a statistical test… and we obtain a p-value.

- If the p-value is below the significance level (for example, 0.05), we can reject the null hypothesis in favor of the alternative hypothesis, i.e. we conclude that the observed difference is the result of a real effect.
- Example: You study some bacteria and it appears to you that their colonies don't live more than 65 days.
- But you want to check if it is true.
- And you want to test your hypothesis at the significance level of 0.05.
- So you take a sample of 157 colonies with known life span for each.
- You calculate the mean (65.12 days) and the standard deviation (9).
- H0: life spam mean = 65
- HA: life spam mean > 65
- It is well known that to test such a hypothesis on the mean of a population there is the z-test.
- The test statistic for this z-test is calculated as (each test uses its own formula to calculate the test statistic): z = (65.12 - 65) / (9 * sqrt(157)) = 0.167

Each test statistics has a known distribution:



which allows to calculate the p-value, that is, to find the probability of observing a test statistic at least this extreme when assuming the null hypothesis.

In our case the p-value to obtain the values of z-test statistic greater than 0.167 is equal 0.3936.

Since this is greater than our significance level, 0.05, we fail to reject the null hypothesis (we are NOT in the green zone of the distribution above).

This means that the data does not support the claim that the mean is greater than 65.

Errors can happen in hypothesis testing:

	Null hypothesis is TRUE	Null hypothesis is FALSE
Reject null hypothesis	Type I error (False positive)	Correct outcome (True positive)
Fail to reject null hypothesis	Correct outcome (True negative)	Type II error (False negative)

Type II errors (false negatives): you are missing some real changes!

Type I errors (false positives): some changes appear to be the result of a real effect while they are not!

Type I errors (false positives) are the most dangerous as they can lead to wrong conclusions.

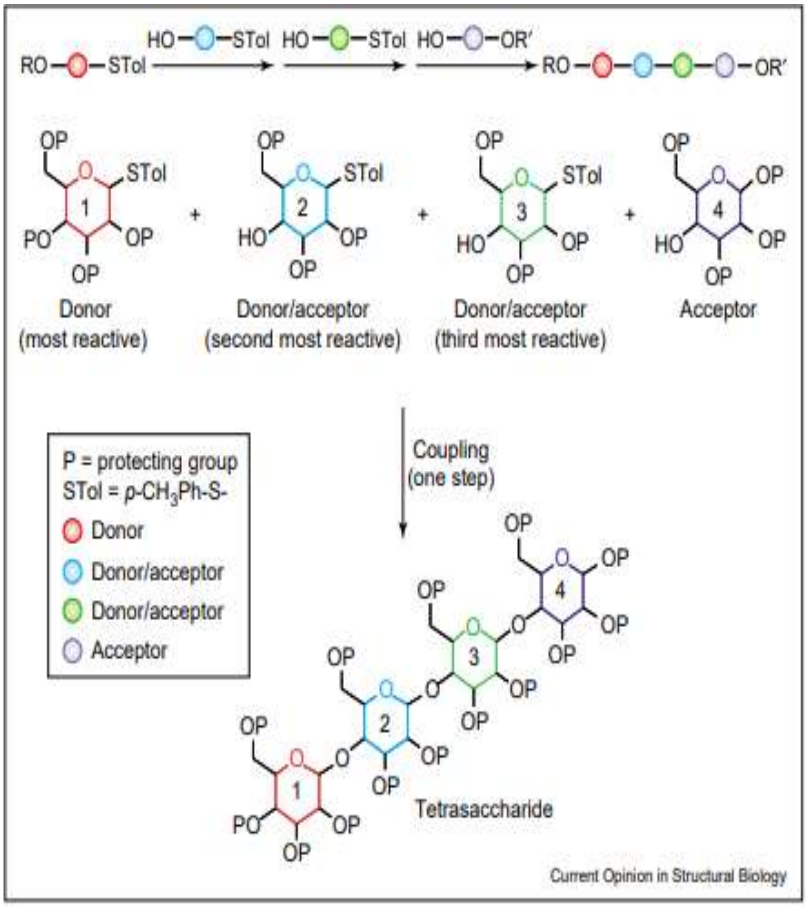
UNIT – 4- SBIA5304-MICROARRAY DATAANALYSIS

UNIT IV CARBOHYDRATE MICROARRAYS

Carbohydrate microarrays- Carbohydrate sources- Synthesis of oligosaccharides - Isolation of oligosaccharides from natural sources- Arrays of monosaccharides and disaccharides- Arrays of polysaccharides- Arrays of oligosaccharides- immunological applications.

- Carbohydrate microarray technologies are new developments at the frontiers of glycomics.
- Results of 'proof of concept' experiments with carbohydrate-binding proteins of the immune system — antibodies, selectins, a cytokine and a chemokine and several plant lectins indicate that microarrays of carbohydrates (glycoconjugates, oligosaccharides and monosaccharides) will greatly facilitate not only surveys of proteins for carbohydrate-binding activities but also elucidation of their ligands.
- It is predicted that both naturally occurring and synthetic carbohydrates will be required for the fabrication of microarrays that are sufficiently comprehensive and representative of entire glycomes.
- New leads to biological pathways that involve carbohydrate–protein interactions and new therapeutic targets are among biomedically important outcomes anticipated from applications of carbohydrate microarrays.
- Unlike proteins and nucleic acids, oligosaccharides are difficult to synthesize chemically.
- This is because some oligosaccharide chains are linear, others are branched, the monosaccharide building blocks are in alpha or beta anomeric configurations, and adjacent monosaccharides are linked via different carbon atoms in their sugar rings.
- For these reasons, multiple selective protection and deprotection steps are required for the hydroxyl groups of monosaccharides during chemical synthesis of oligosaccharides; the manual synthesis of oligosaccharides is a major undertaking
- The solid-phase synthesis approach has the advantage of avoiding intermediate isolation and purification steps.
- An automated solid-phase method that includes selective protection and deprotection steps has been introduced and applied to the synthesis of several glucose- and mannose-containing oligosaccharides

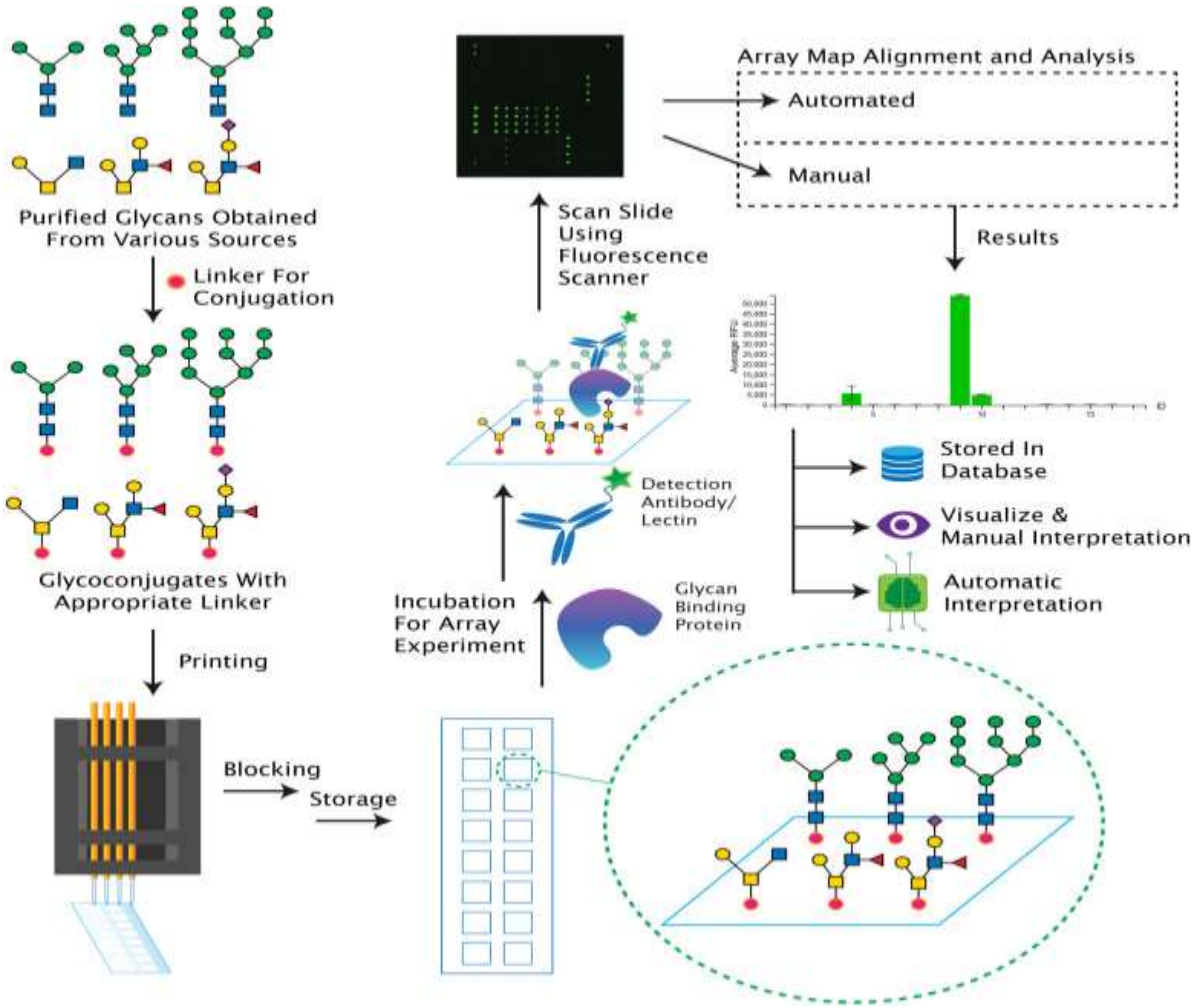
- An alternative approach to the synthesis of oligosaccharides is a programmable ‘one-pot’ approach, in which an oligosaccharide of interest is generated by the sequential addition of building blocks (thioglycosides) that are either fully protected or have one hydroxyl group exposed.



Schematic representation of a programmable one-pot approach to oligosaccharide synthesis; a linear 1,4-linked tetrasaccharide is depicted as an example. The constituent building blocks are classified into three species: the first sugar at the nonreducing end acts as the donor; the last sugar at the reducing end is the acceptor; all other building blocks that form the inner part of a complex (linear or branched) oligosaccharide are classified as donor/acceptor. Protecting groups (esters or ethers) determine the RRV of anomeric centers. Building blocks are added in the order 1 to 4 to obtain the tetrasaccharide.

- It has been shown that the relative reactivity value (RRV) of a thioglycoside building block in the glycosidation reaction can be tuned in the presence of protecting groups; more than 200 building blocks, with RRVs ranging from 1 to 105 , have been designed and synthesized.
- A computer programme called ‘Optimer’ has been developed to guide the selection of building blocks for the one-pot synthesis of a given oligosaccharide.
- If RRVs differ by more than 102 , the desired glycosidic bonds will be formed by the sequential addition of building blocks in the order of the RRV values.

- Once the required building blocks with protecting groups are prepared, oligosaccharides can be synthesized in a short period of time (in minutes or hours, instead of days or months using traditional methods) using this programmable one-pot approach
- Glycans are one of the major biological polymers found in the mammalian body.
- They play a vital role in a number of physiologic and pathologic conditions.
- Glycan microarrays allow a plethora of information to be obtained on protein–glycan binding interactions.
- Overview of a typical glycan microarray workflow

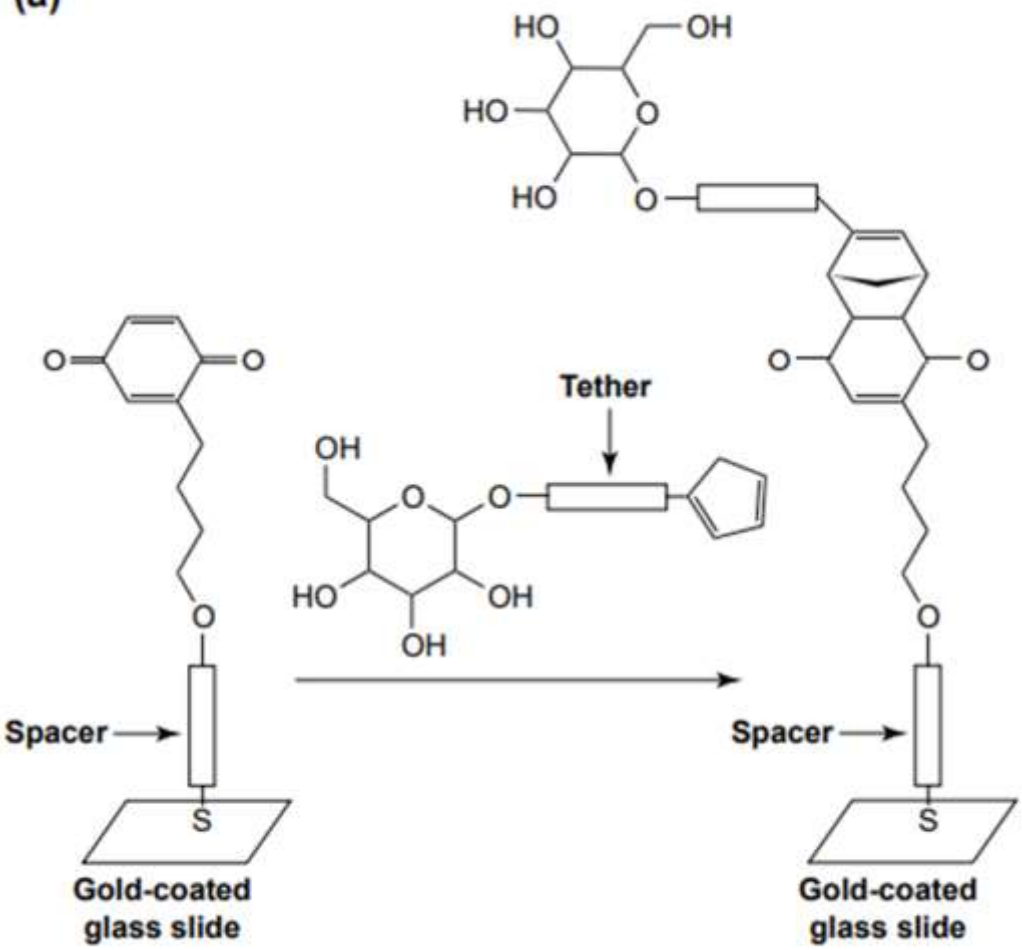


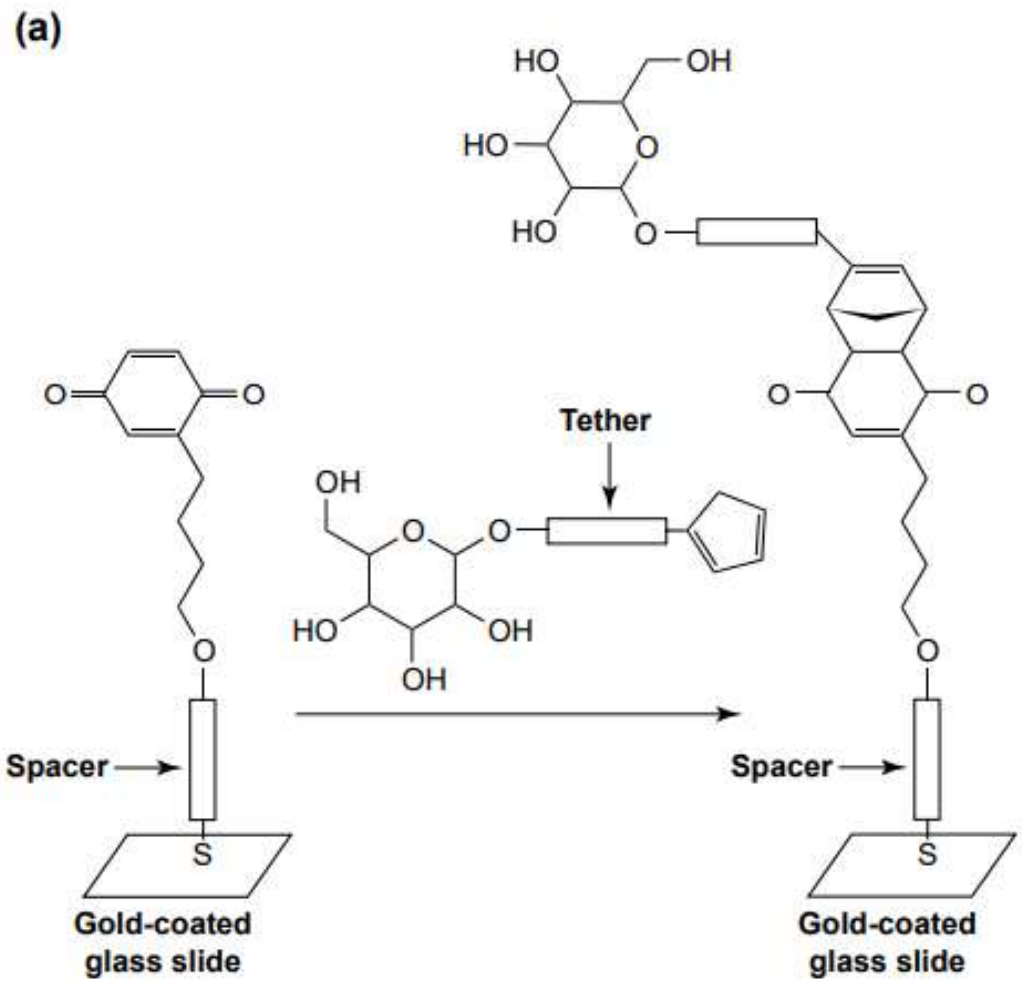
- **Glycans** are chemically or enzymatically synthesized, or isolated and purified from either source materials, and then conjugated with a linker which is appropriate for the printing surface.

- The **glycoconjugates** are then printed upon appropriately functionalized slides, followed by blocking; the printed slides are stored under ideal conditions prior to experiments.
- Many **arrays can be printed on a single slide**, termed sub-arrays. The slides can then be used in a glycan microarray experiment where they are incubated with a **glycan binding protein (GBP),** such as lectin, antibody, or serum, virus, etc., Followed by addition of a **detection reagent**, if the primary analyte was not fluorescently labeled, for example a fluorescent secondary antibody or streptavidin.
- After washing the slide to remove unbound material, the bound material is then identified and measured by scanning using a fluorescence microarray scanner. The image produced can then be analyzed using automated or manual methods to generate the array results. These results can in turn be stored in a database, or interpreted either manually or by automatic algorithms.
- Isolation of oligosaccharides from natural sources
- Oligosaccharides with reducing termini are ideal for derivatization so that they can be immobilized.
- Free reducing oligosaccharides may be isolated from human or animal milk and urine, or they may be in the form of N-linked glycoprotein oligosaccharides released by the enzymes peptide-N-(N-acetyl-b-glucosaminyl)asparagine amidase (PNGase F) and endo-b-N-acetylglucosaminidase F (Endo F) or by hydrazinolysis.
- O-linked glycoprotein oligosaccharides may be released by mild alkaline hydrolysis or hydrazinolysis.
- Oligosaccharides may, if desired, be released from glycolipids by endoceramidase .
- Oligosaccharide fragments can be obtained from proteoglycans and glycosaminoglycans by lyase digestion or nitrous acid degradation , and, in the case of hyaluronic acid, also by hydrolase digestion .
- Various chemical methods may be used to obtain oligosaccharide fragments from bacterial and plant polysaccharides; these include acid or alkaline hydrolysis, acetolysis and Smith degradation.
- Reduced oligosaccharides (oligosaccharide alditols) can be manipulated chemically at the reduced end after mild periodate oxidation to cleave the terminal open chain monosaccharide residue and create a reactive aldehyde for derivatization.

- Reduced oligosaccharides are typically obtained when O-linked glycans are released from glycoproteins by reductive alkaline hydrolysis.
- Oligosaccharide alditols are also available when reduction is carried out, for example, before HPLC separation, to eliminate double peaks resulting from the resolution of a and b anomers at their reducing ends.
- Multiple chromatographic steps are often necessary for the isolation/purification of oligosaccharides.
- These include gel filtration, weak and strong anion-exchange chromatography, thin-layer chromatography (TLC),normal-phase HPLC with an amine or amide column, and reversed-phase HPLC using a C18 or graphitized porous carbon column.
- **Arrays of monosaccharides and disaccharides**
- The monosaccharides were covalently immobilized by conjugation to self-assembling monolayers of alkenethiols on the gold surface.
- The first step was to prepare monolayers consisting of two alkenethiols, one of which has a benzquinone group exposed.
- The monosaccharides, in the form of diene conjugates, are then applied as 1 ml spots (2 mM in water) and attached to the slides through the Diels–Alder cycloaddition reaction.
- This is a very high yielding process, often reagent free, and moisture and solvent tolerant, and is therefore ideal for the microarray format

(a)


HO
OH
HO
O
HO
O
O
O
O
O
Tether
OH
O
O
HO
OH
OH
O
Spacer
S
Gold-coated
glass slide
Spacer
S
Gold-coated
glass slide



- The covalently immobilized monosaccharides were evaluated by profiling the binding specificities of five plant lectins, ***concanavalin A (Con A), Benderia simplicifolia, Erythrinia cristalli, Ulex europeeus and Galanthus nivalis***, that are known to bind to different monosaccharides.
- Specific monosaccharide binding was observed for the five lectins, which were labeled fluorescently with rhodamine. Specific binding of Con A to arrayed mannose was also shown by surface plasmon resonance spectroscopy.
- In further experiments, the monosaccharide array was probed with the glycosyltransferase b-1,4-galactosyltransferase; it was shown that enzyme-mediated glycosylation of immobilized N-acetylglucosamine occurred in the presence of the donor substrate, UDP-galactose.

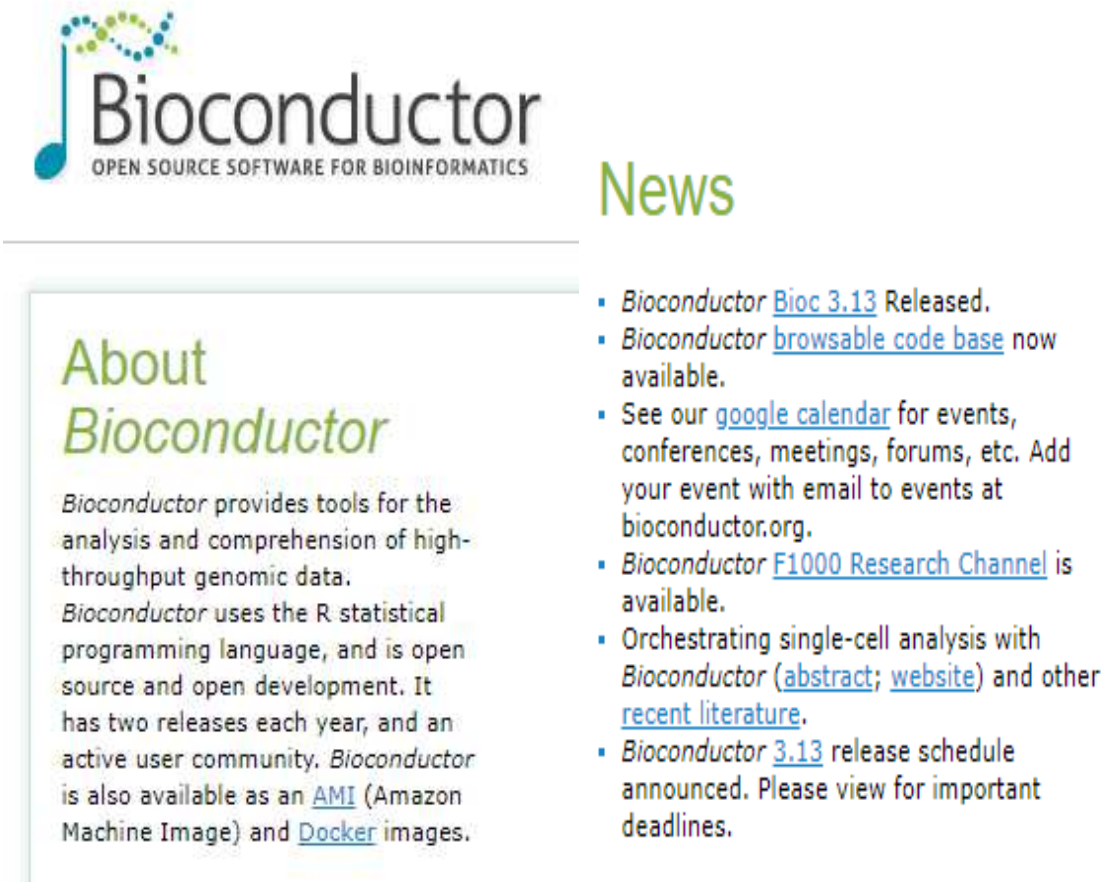
- Shin’s group has reported another approach to carbohydrate microarray fabrication. They used glass slides modified by thiol groups as solid supports.
- One monosaccharide, N-acetylglucosamine, and three disaccharides, lactose, cellobiose and maltose, in the form of glycosylamines, were converted into maleimide conjugates and then covalently bound to the glass surface by hetero-Michael addition reaction between the thiol group on the solid surface and the maleimide moiety of the sugar derivative .
- The maleimide-conjugated carbohydrates (from 0.1 to 5.0 mM) were printed with a pin-type microarrayer on the slides at a spot size of 100 mm and a pitch of 200 mm. Carbohydrate–protein interaction studies were performed with fluoresce in labeled plant lectins.
- The binding of the three plant lectins examined, Con A, Erythrina cristagalli and Triticum vulgaris, to the monosaccharide and the disaccharides was in accord with their known specificities.
- **Arrays of polysaccharides**
- Wang et al. described microarrays of polysaccharides and glycoproteins on nitrocellulose-coated glass slides.
- They used a high-precision robotic arrayer that was developed for cDNA and the spots were generated without derivatization.
- The spot sizes were 150 mm with a pitch of 375 mm. These were air dried to allow adsorption (noncovalent immobilization) onto the hydrophobic surface.

UNIT – 5- SBIA5304-MICROARRAY DATAANALYSIS

UNIT V DATABASES AND TOOLS FOR MICROARRAYS

- Bioinformatics in Arrays- Databases and tools for microarrays- Bioconductor, expression profiler, EST databases- Assessing levels of gene expression using EST's, TIGR gene indices, STACK, SAGE, CGAP, Xprofiler, ARRAY DB, cluster, tree view, Scanalyze, gene cluster, informatics aspects of microarray production- MGED and gene-ontology, description of MIAME ((Minimum Information About a Microarray Experiment), Business Aspects of Biochip Technologies- Microarray Technology in Treating Disease.

Bioconductor



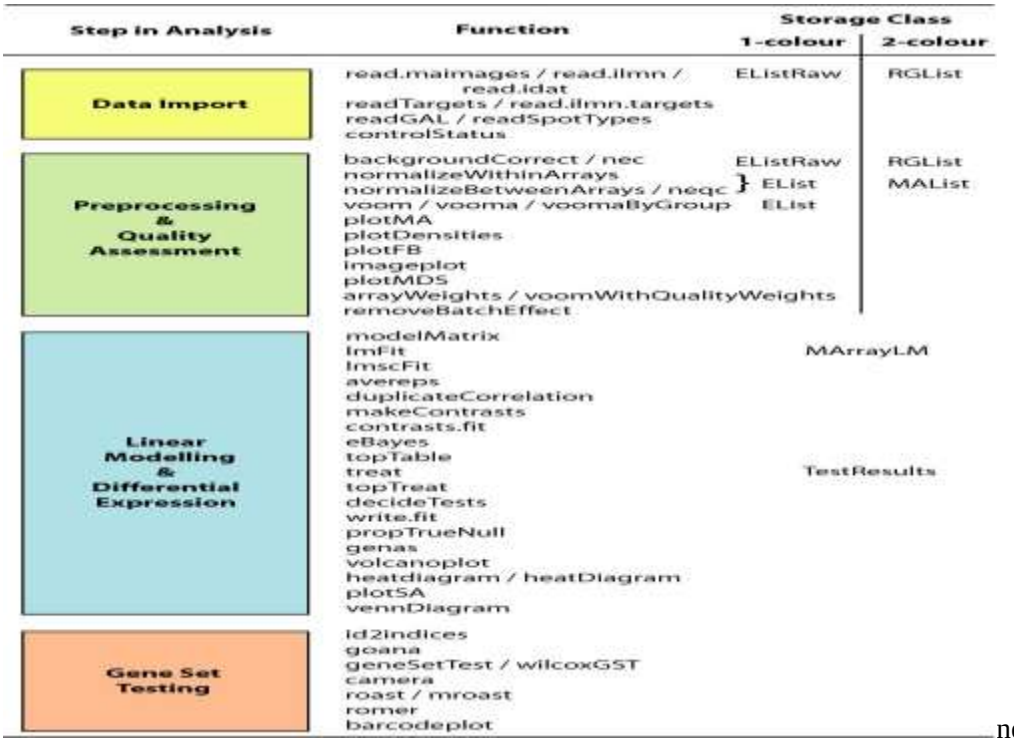
What is Bioconductor used for?

- Bioconductor provides tools for the analysis and comprehension of high-throughput genomic data.
- Bioconductor uses the R statistical programming language, and is open source and open development.

- It has two releases each year, and an active user community.

limma powers differential expression analyses for RNA-sequencing and microarray studies

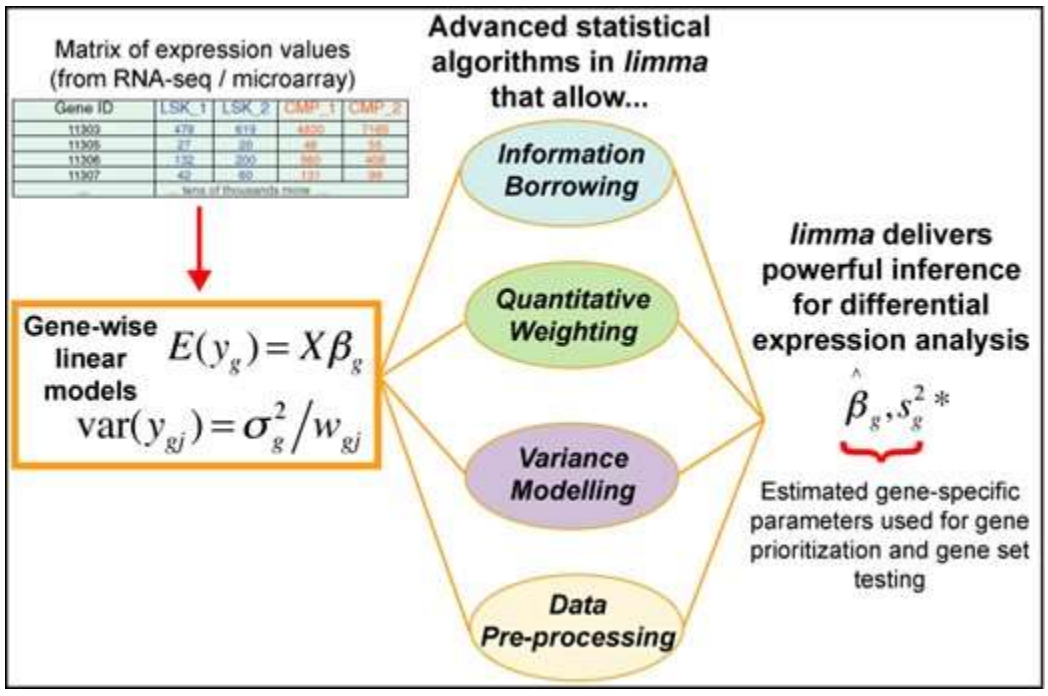
limma is an R/Bioconductor software package that provides an integrated solution for analysing data from gene expression experiments. It contains rich features for handling complex experimental designs and for information borrowing to overcome the problem of small sample sizes. Over the past decade, limma has been a popular choice for gene discovery through differential expression analyses of microarray and high-throughput PCR data. The package contains particularly strong facilities for reading, normalizing and exploring such data. Recently, the capabilities of limma have been significantly expanded in two important directions. First, the package can now perform both differential expression and differential splicing analyses of RNA sequencing (RNA-seq) data. All the downstream analysis tools previously restricted to microarray data are now available for RNA-seq as well. These capabilities allow users to analyze both RNA-seq and microarray data with very similar pipelines. Second, the package is now able to go past the traditional gene-wise expression analyses in a variety of ways, analyzing expression profiles in terms of co-regulated sets of genes or in terms of higher-order expression signatures. This provides enhanced possibilities for biological interpretation of gene expression differences.



ne

The limma package is a core component of Bioconductor, an R-based open-source software development project in statistical genomics. It has proven a popular choice for the analysis of data from experiments involving microarrays, high-throughput polymerase chain reaction (PCR), protein arrays and other platforms. The package is designed in such a way that, after initial pre-processing and normalization, the same analysis pipeline is used for data from all technologies.

Recently, the capabilities of limma have expanded significantly in two important directions. First, the package can now perform both differential expression (DE) and differential splicing analyses of RNA sequencing (RNA-seq) data. All the downstream analysis tools previously restricted to microarray data are now available for RNA-seq as well. These capabilities allow users to analyse both RNA-seq and microarray data with very similar pipelines. Second, the package is now able to go past the traditional gene-wise expression analyses in a variety of ways, analysing expression profiles in terms of co-regulated sets of genes or in terms of higher-order expression signatures. This provides enhanced possibilities for biological interpretation of gene expression differences.



EST databases- Assessing levels of gene expression using EST's

- EST expression profiling has by now become well-established high-throughput method for acquiring quantitative information on a sample's transcriptome and for studying differential gene expression, inferred from the differences in the relative numbers of EST tags between two libraries.

- To facilitate gene discovery, the EST content of a library can be altered to reduce the abundance of transcripts representing genes with high expression.
- To achieve this a library can be either normalised by removing the most abundant transcripts in order to reduce or eliminate the differences in the relative transcript abundances to a narrow range , or subtracted to enrich the library for rare novel transcripts .
- Ideally this should create a library containing the same or similar tag counts for the low abundance sequences as before, but with vastly reduced counts for abundant or unwanted cDNAs.
- Neither normalised nor subtracted libraries are suitable for studying differential mRNA expression because of the significantly changed representation or removal of the original transcripts

The TIGR Gene Indices

- The TIGR Gene Indices (http://www.tigr.org/tdb/tgi) are a collection of 77 species-specific databases that use a highly refined protocol to analyze gene and EST sequences in an attempt to identify and characterize expressed transcripts and to present them on the Web in a user-friendly, consistent fashion.
- A Gene Index database is constructed for each selected organism by first clustering, then assembling EST and annotated cDNA and gene sequences from GenBank. This process produces a set of unique, high-fidelity virtual transcripts, or tentative consensus (TC) sequences. The TC sequences can be used to provide putative genes with functional annotation, to link the transcripts to genetic and physical maps, to provide links to orthologous and paralogous genes, and as a resource for comparative and functional genomic analysis.

Summary of the current release of TIGR Gene Indices (TGI)

Species	Species_name	TGI	TC	sET	sEST
Animals (29)					
Human	*Homo sapiens*	HGI 15.0	221 418	19 740	594 468
Mouse	*Mus musculus*	MGI 14.0	167 694	7499	602 312
Rat	*Rattus norvegicus*	RGI 13.0	56 933	2131	87 992
Cattle	*Bos Taurus*	BtGI 10.0	38 760	413	56 644
Pig	*Sus scrofa*	SsGI 9.0	33 963	519	50 376
Dog	*Canis familiaris*	DogGI 4.0	6613	684	11 506
Chicken	*Gallus gallus*	GgGI 8.0	42 988	848	72 941
Frog	*Xenopus laevis*	XGI 9.0	39 724	626	37 249
Zebrafish	*Danio rerio*	ZGI 15.0	32 889	395	53 940
Catfish	*Ictalurus punctatus*	Cfgi 5.0	3254	156	16 694
R.trout	*Oncorhynchus mykiss*	RtGI 4.0	23 135	190	27 448

Construction of the Gene Indices

The process used to assemble each Gene Index is similar to that described previously, although some modifications have been made to improve the efficiency and accuracy of the process. mgBLAST, a modified version of the Megablast program, is now used for the pairwise sequence comparisons that are the basis for defining the sequence clusters which form the basis for assembly. For large clusters containing hundreds or thousands of sequences (e.g. highly expressed genes such as actin), sequence representation is reduced prior to assembly using a variety of multilayer approaches, including transitive clustering, containment clustering and seeded clustering with known genes. Following clustering, the Paracel Transcript Assembler (PTA), a modified version of CAP3 assembly program, is used to assemble each TC. An open source set of software tools that embody this process, TGICL, is available (http://www.tigr.org/tdb/tgi/software) with other open-source utilities for users interested in performing a similar analysis on their own datasets.

SAGE

Serial Analysis of Gene Expression (SAGE) is a transcriptomic technique used by molecular biologists to produce a snapshot of the messenger RNA population in a sample of interest in the form of small tags that correspond to fragments of those transcripts.

Serial analysis of gene expression (SAGE) uses mRNA from a particular sample to create complementary DNA (cDNA) fragments which are then amplified and sequenced using high-throughput sequencing technology.

The mechanism behind SAGE is based on tags which can identify the original transcript, and rapid sequencing of chains of tags linked together. The procedure essentially simplifies sequencing by linking the cDNA segments together in a long chain.

The resulting analysis gives a snapshot of the transcriptome of the sample, including the identity and abundance of each mRNA.

Steps of SAGE

SAGE is a complex protocol with many steps.

Step 1: mRNA is isolated from the sample and reverse transcribed using biotinylated primers to generate cDNA

Step 2: cDNA is bound via biotin to streptavidin microbeads

Step 3: cDNA is cleaved with restriction enzymes freeing it from the beads

Step 4: Cleaved DNA is washed out, leaving truncated cDNA bound to the beads

Step 5: Two oligonucleotides with sticky ends are added to the remaining truncated cDNA, in separate samples

Step 6: Cleaved DNA is “tagged” enzymatically, removing it from the beads

Step 7: Sticky ends are repaired with DNA polymerase

Step 8: Blunt ended tags from the two separate samples are ligated together, generating ditags with two different oligonucleotide adapter ends

Step 9: Ditags are cleaved to remove the oligonucleotides. Ditags will form long cDNA chains, or concatemers

Step 10: Transform concatemers into bacteria for replication

Step 11: Isolate concatemers from bacteria and sequence

Challenges when using SAGE

One challenge is that the tags are only about 13 or 14 base pairs. It can be difficult to identify such a short tag if it’s from an unknown gene.

The flip side of that problem is that SAGE can be used to find unknown genes, and in some studies it’s an advantage to be able to measure gene expression quantitatively without prior sequence information.

Tags may also have issues with specificity; multiple genes could share the same tag if there is an overlap in sequence. There also can be inconsistencies with the restriction enzymes, and incompatibilities for certain species.

SAGE and DNA microarray

SAGE is similar in many ways to a DNA microarray; however, in a DNA microarray, the mRNAs hybridize to cDNA probes on the array. In SAGE, the data output is based on sequencing. That means SAGE analysis is more quantitative and it does not depend on the use of known genes.

Microarray experiments are generally less costly, and so are used more often in larger-scale studies.

Application

- A study of new markers in cancer illustrates how SAGE can be used in biomedical research.
- Researchers compared gene expression levels in cancerous tissues with those in non-cancerous tissues to search for markers that could diagnose the pancreatic cancer at an early stage.
- Because the results of a SAGE analysis of many representative tissues had already been published online, the scientists were able to search the database for genes preferentially expressed in pancreatic cancer.
- From this, they were able to identify a gene calledprostate stem cell antigen (PCSA), that had previously not been associated with pancreatic cancer.

- **CGAP- Cancer Genome Anatomy Project**

- **What you can do:**

- Find the information and technological tools needed to decipher the molecular anatomy of the cancer cell from a annotated index of the genes that are important in cancer.

- **Highlights:**

- The goal of the NCI's Cancer Genome Anatomy Project is to determine the gene expression profiles of normal, precancer, and cancer cells, leading eventually to improved detection, diagnosis, and treatment for the patient.

- CGAP offers technological, informational (data and analysis tools), resource (clones and libraries) and methodological infrastructure for the cancer research community.

- The current CGAP program has expanded to include in addition to the Tumor Gene Index (TGI), a Genetic Annotation Initiative (GAI) and the Cancer Chromosome Aberration Project (cCAP).

- The TGI and GAI are focused towards building a catalog of annotated genes.